

NYC Uber Demand Forecasting based on spatial-temporal analysis

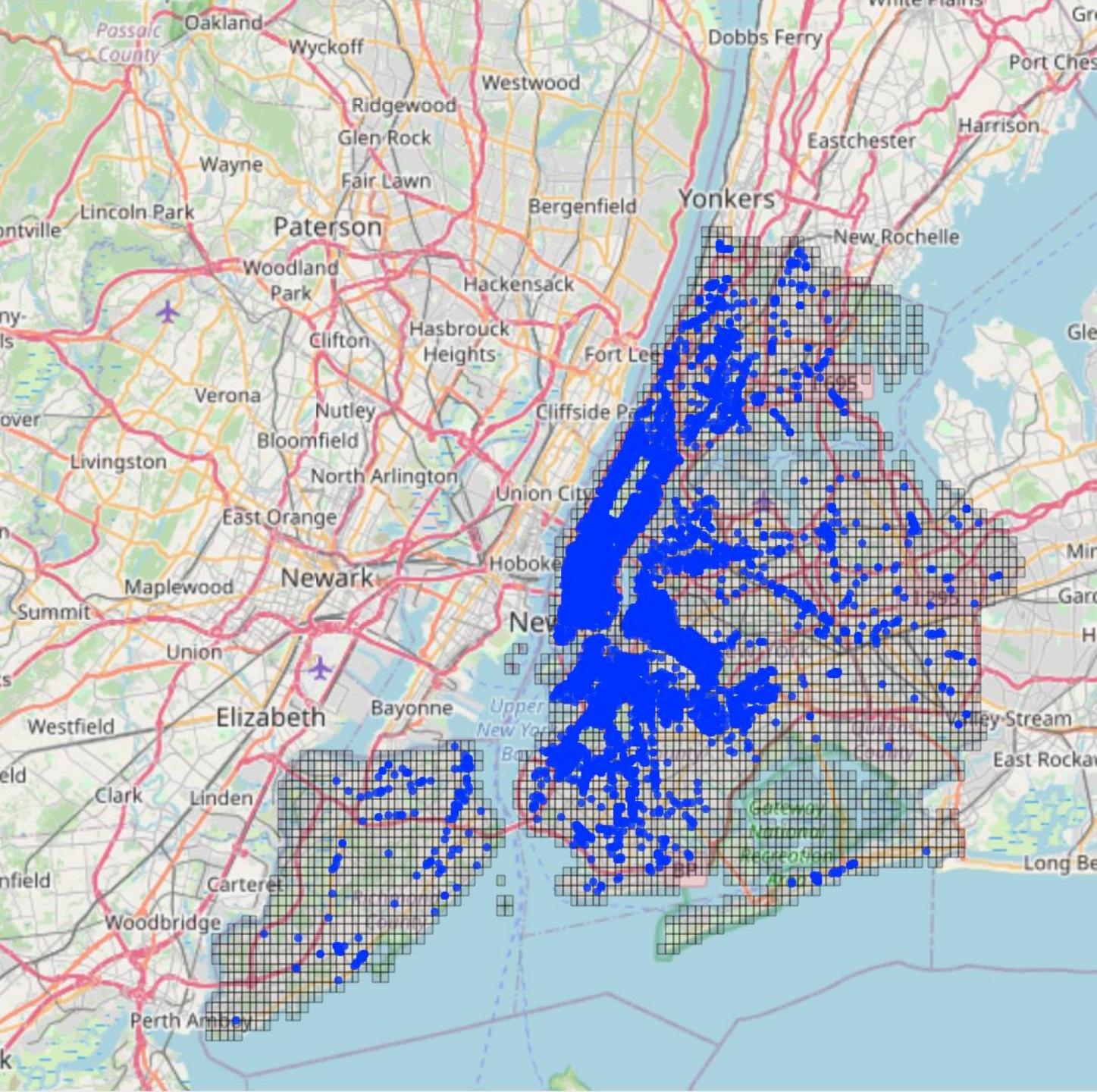
Tingting Huang & Jing Zhang

2020-April

Agenda

- Purpose of the project
- Methodology
- Exploratory Analysis of NYC Uber Data and it's relationship with other variables
- Model Result : Accuracy

1. Purpose



Why Forecast Uber Demand

Uber and other ridesharing ideas have become a popular mode and an important part of the transportation system in recent years. To forecast the Uber demand accurately and efficiently will benefit both **Transportation Planners** and **Uber users (drivers & passengers)**



For Transportation Planners

“Uber took down the **TAXI** industry and now it wants a piece of **PUBLIC TRANSIT**”

--- CNN Business, 2019

The collaboration between ridesharing and traditional transportation modes (e.g. transit & taxi) are limited. Transportation planners need to understand the current/future demand of ridesharing to coordinate the transport system as a whole.



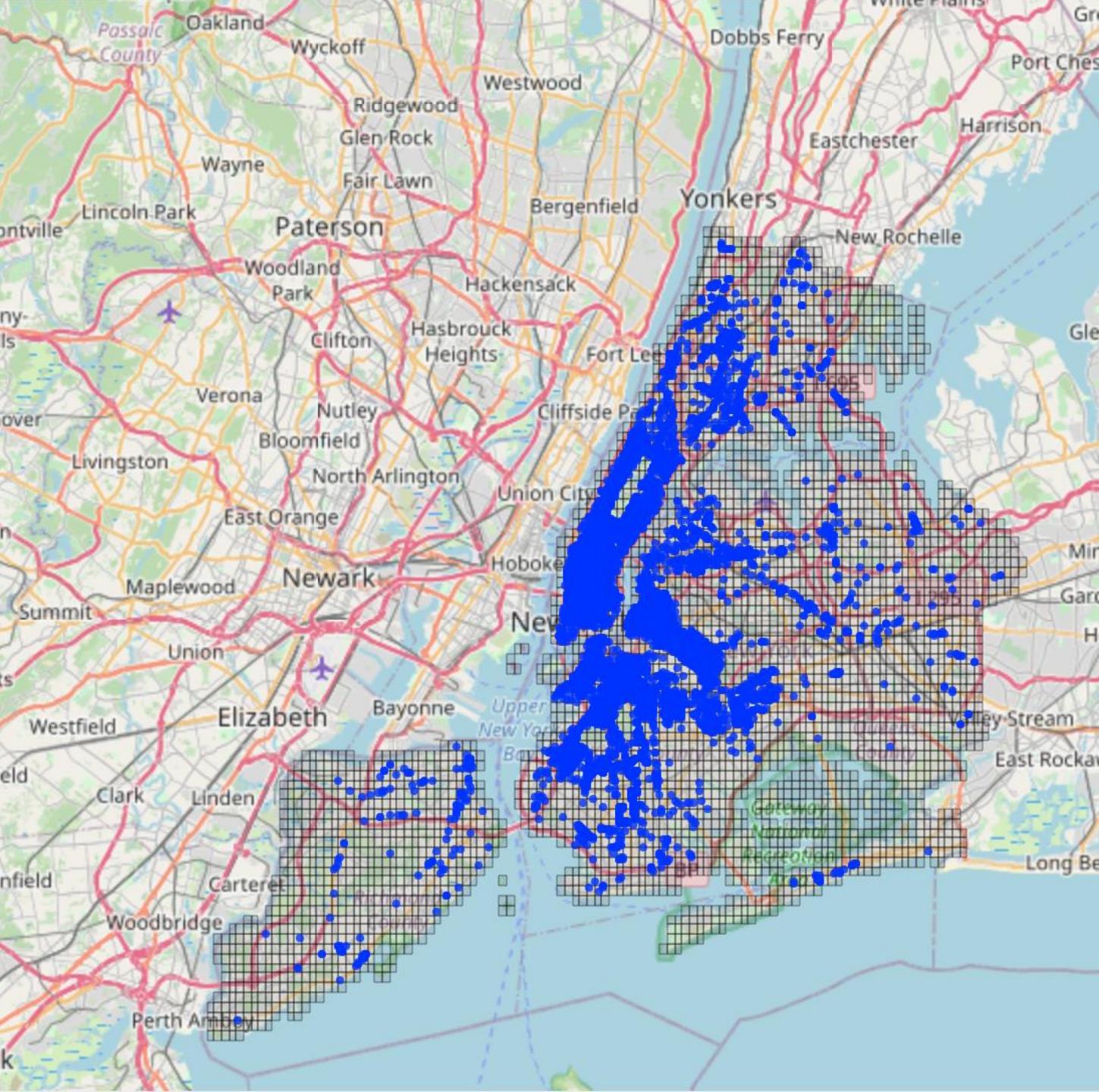
For Uber Users (driver & passengers)

Better user experience;
Inform drivers and passengers about the predicted Uber demand will help them to arrange their trip plan and improve the efficiency.

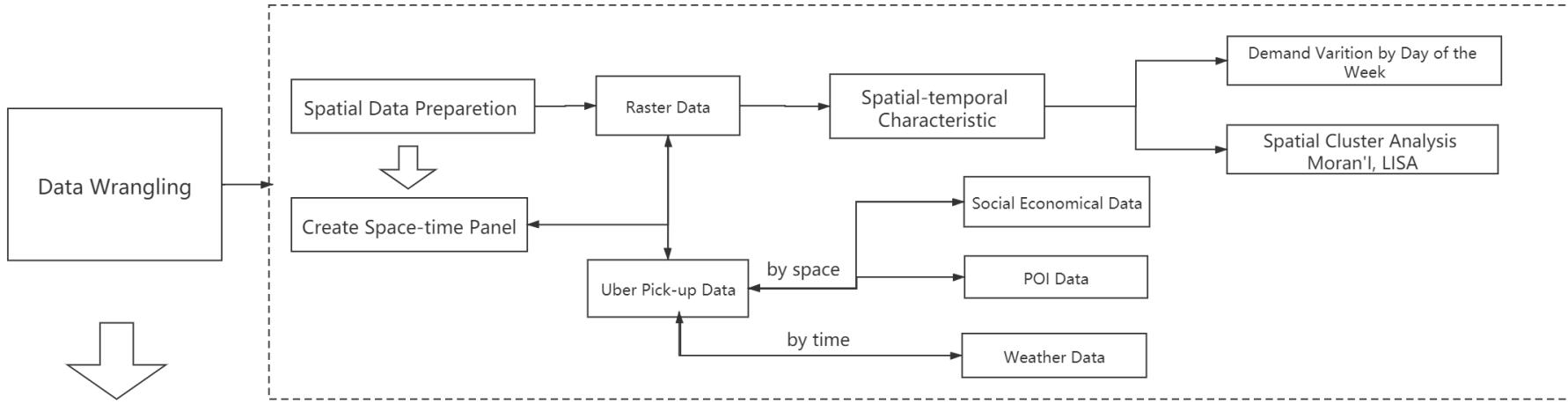
Purposes of the Project

1. Explore the NYC's Uber demand **distribution** in 2014-May
2. Explore which variables **impact** NYC's Uber demand (pickup)
3. Build an **accurate and generalize** predictive model based on Linear regression

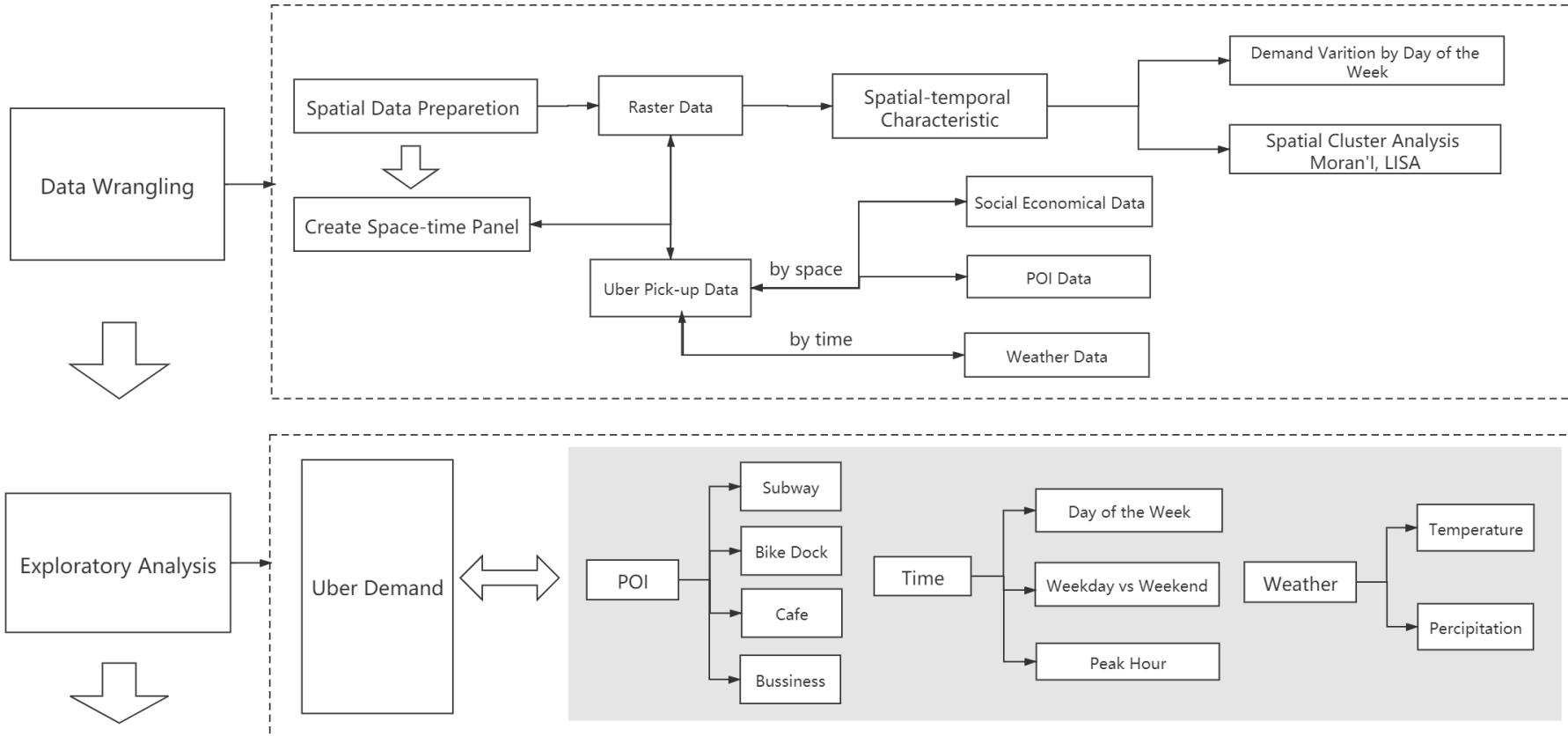
2. Methodology



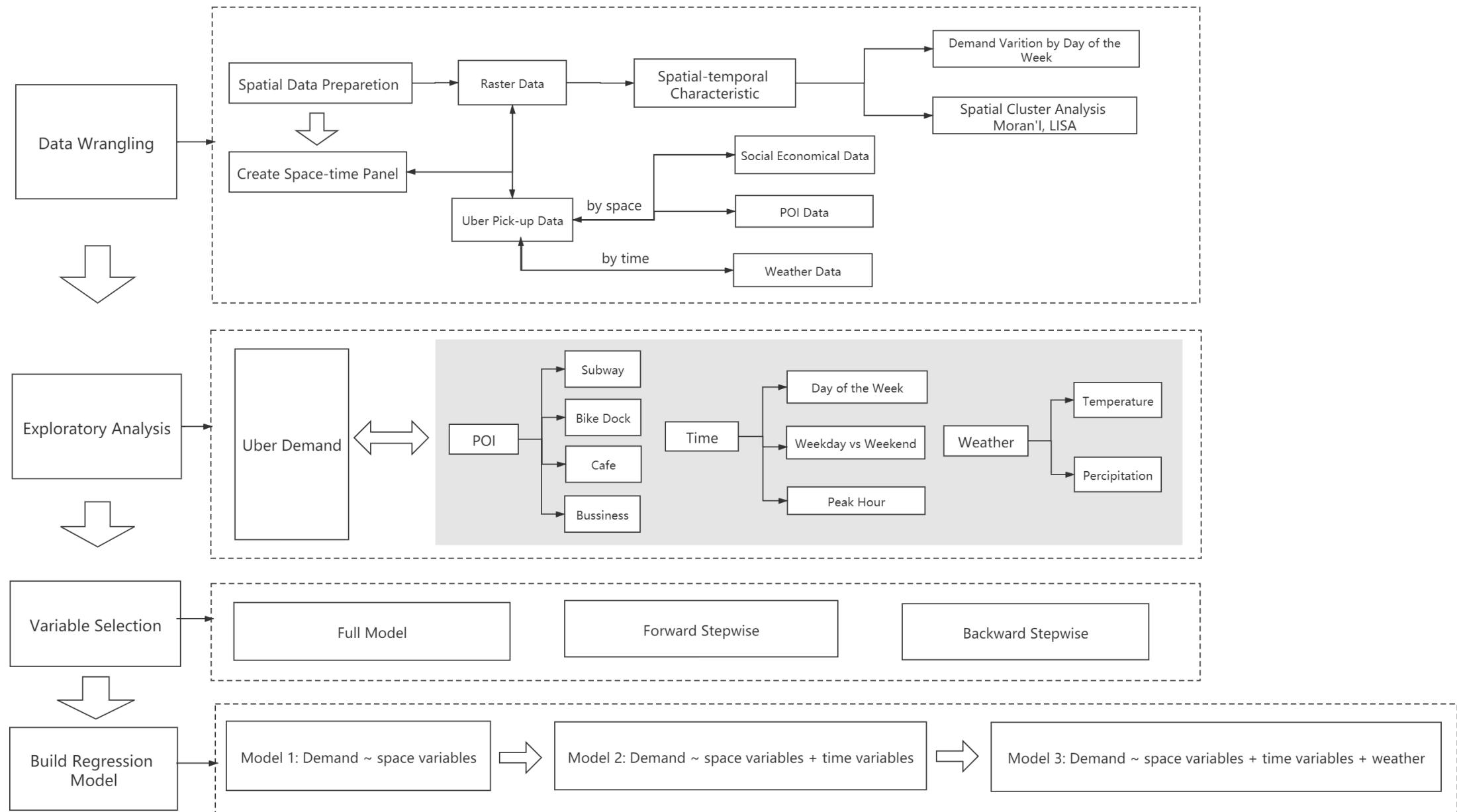
Methodology



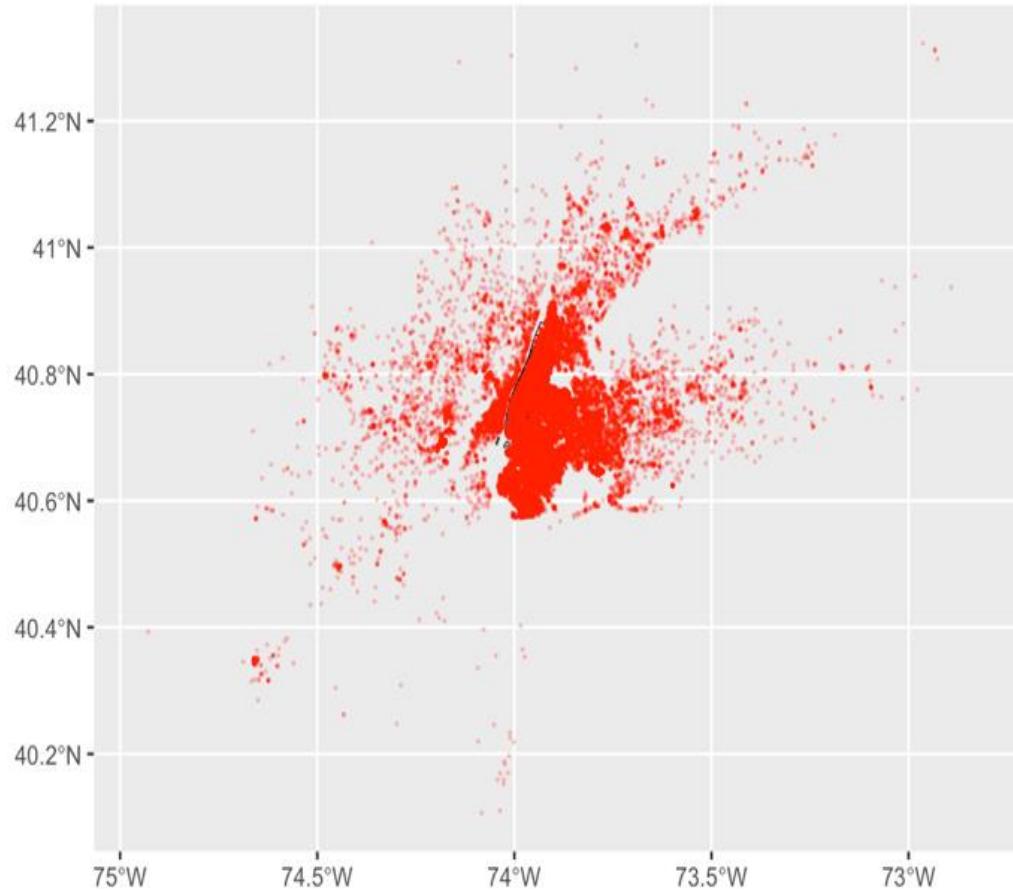
Methodology



Methodology



NYC Uber Pickup data



| | |
|-----------|-----------------------------------|
| Data Name | 2014-May NYC Uber pickup raw data |
| Source | Kaggle |
| Size | 652,435 row |

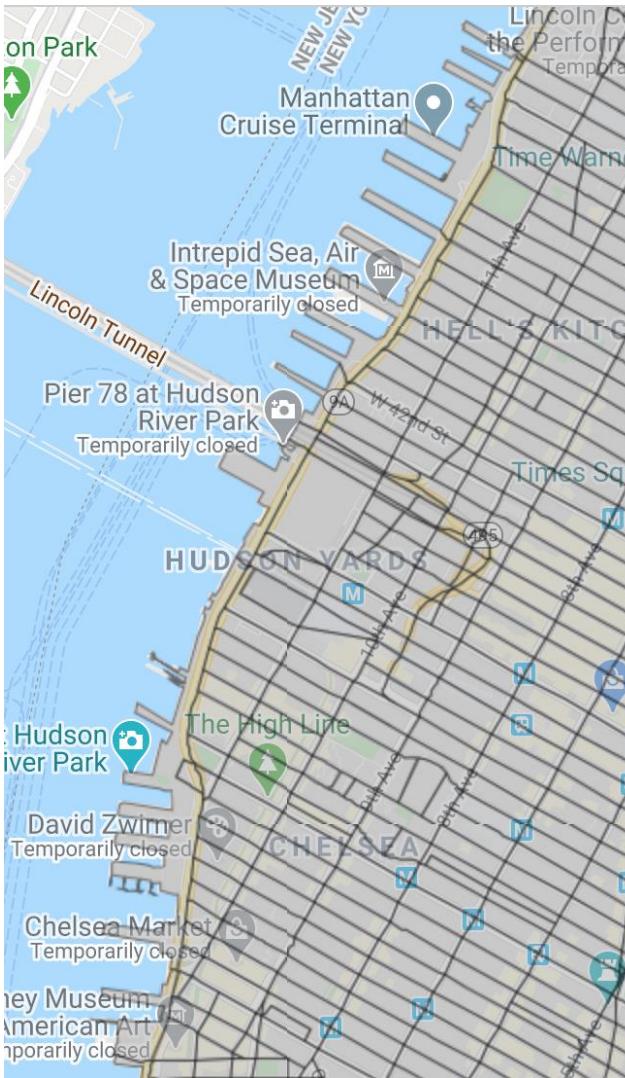
| | Date.Time | Lat | Lon |
|---|------------------|---------|----------|
| 1 | 5/1/2014 0:02:00 | 40.7521 | -73.9914 |
| 2 | 5/1/2014 0:06:00 | 40.6965 | -73.9715 |
| 3 | 5/1/2014 0:15:00 | 40.7464 | -73.9838 |
| 4 | 5/1/2014 0:17:00 | 40.7463 | -74.0011 |
| 5 | 5/1/2014 0:17:00 | 40.7594 | -73.9734 |
| 6 | 5/1/2014 0:20:00 | 40.7685 | -73.8625 |

Analysis Unit - Why Raster

NYC Census Tract



NYC Census Block



Analysis Unit - Why Raster

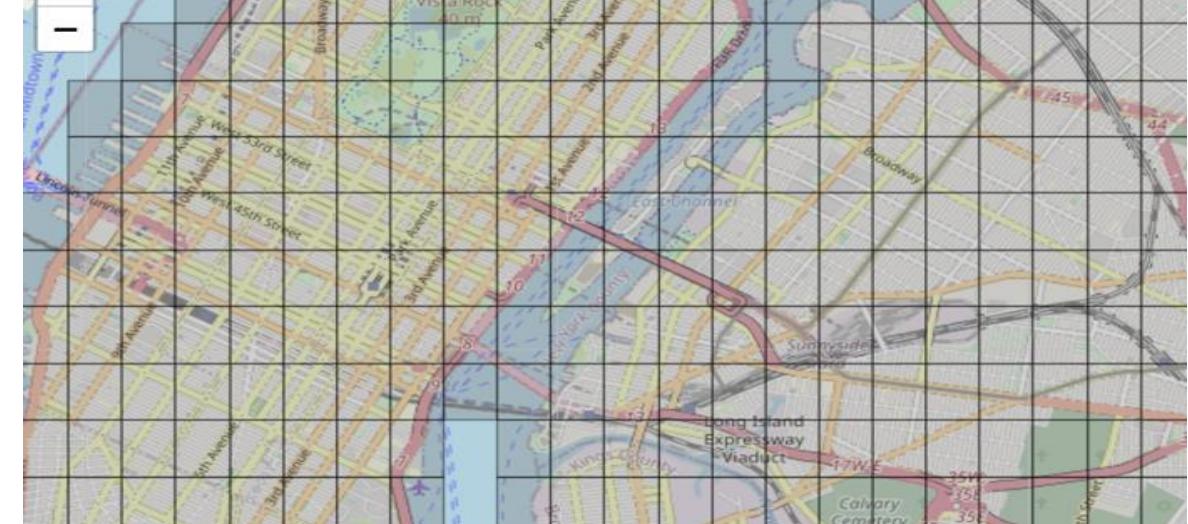
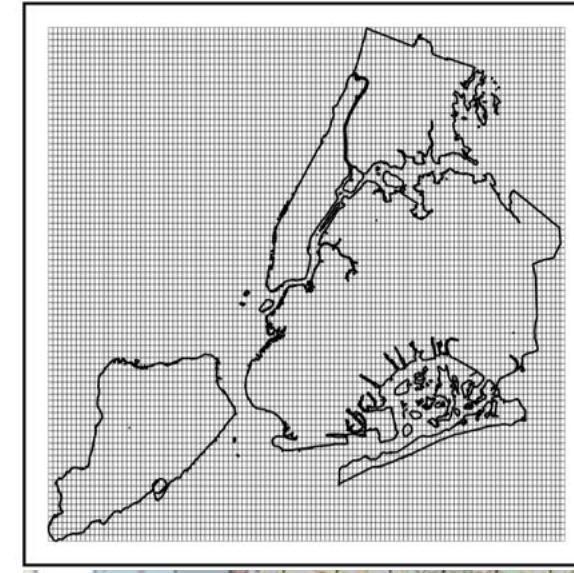
NYC Census Tract



NYC Census Block



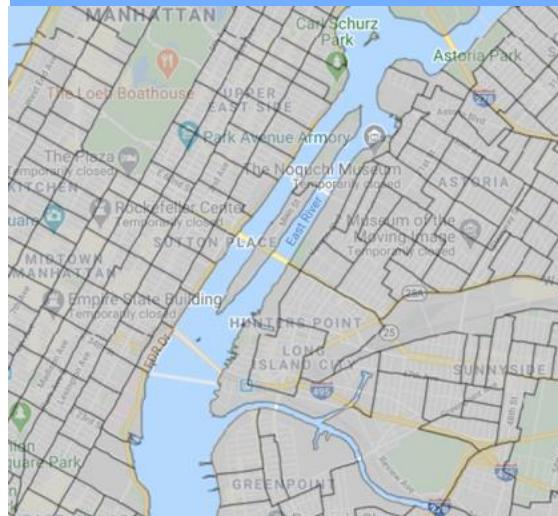
NYC & 500*500m grids



Analysis Unit - Why Raster

NYC Census Tract

The size of census tracts is different, difficult to observe the real demand



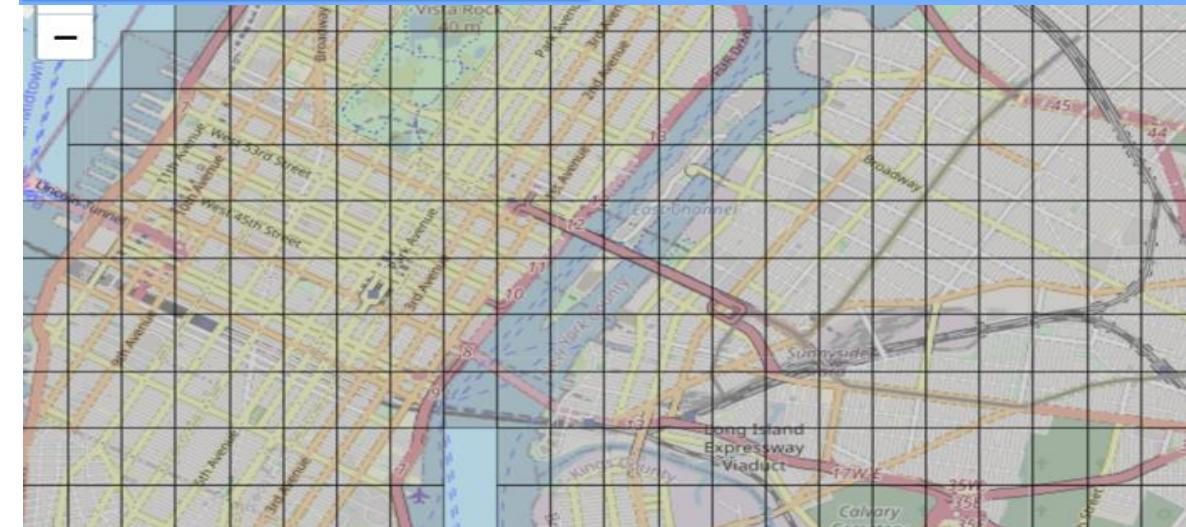
NYC Census Block

Size too small, can't count the pickup within the blocks since the boundary is the street



NYC & 500*500m grids

Same size, easier to find out which small grid is the busiest one, and help the Transportation planners or Uber users to make a decision. Because of the coding limitation, we didn't rotate the grids to make it in line with the street direction, but we should.



Related Variables

Time

Weather

Space

NYC Uber Demand Related Variables



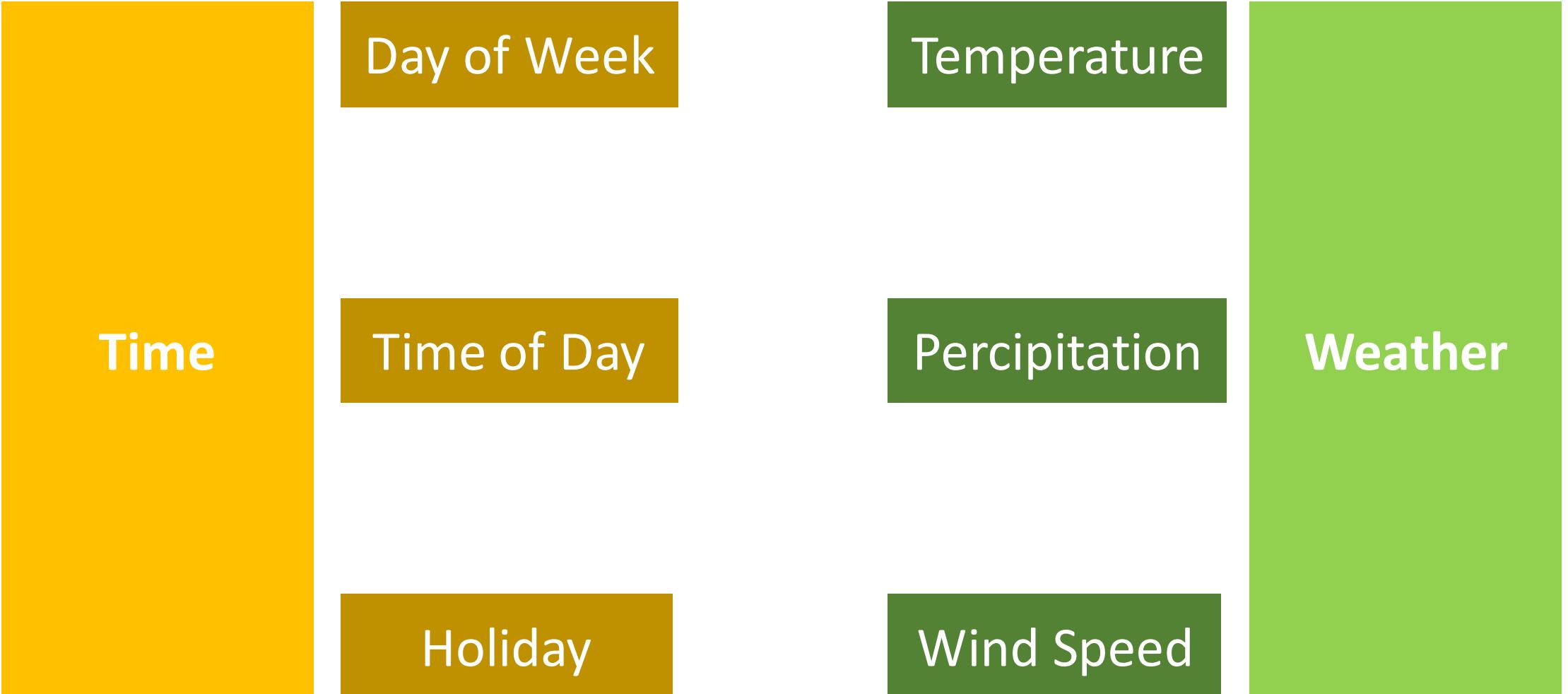
Time of Day

Holiday

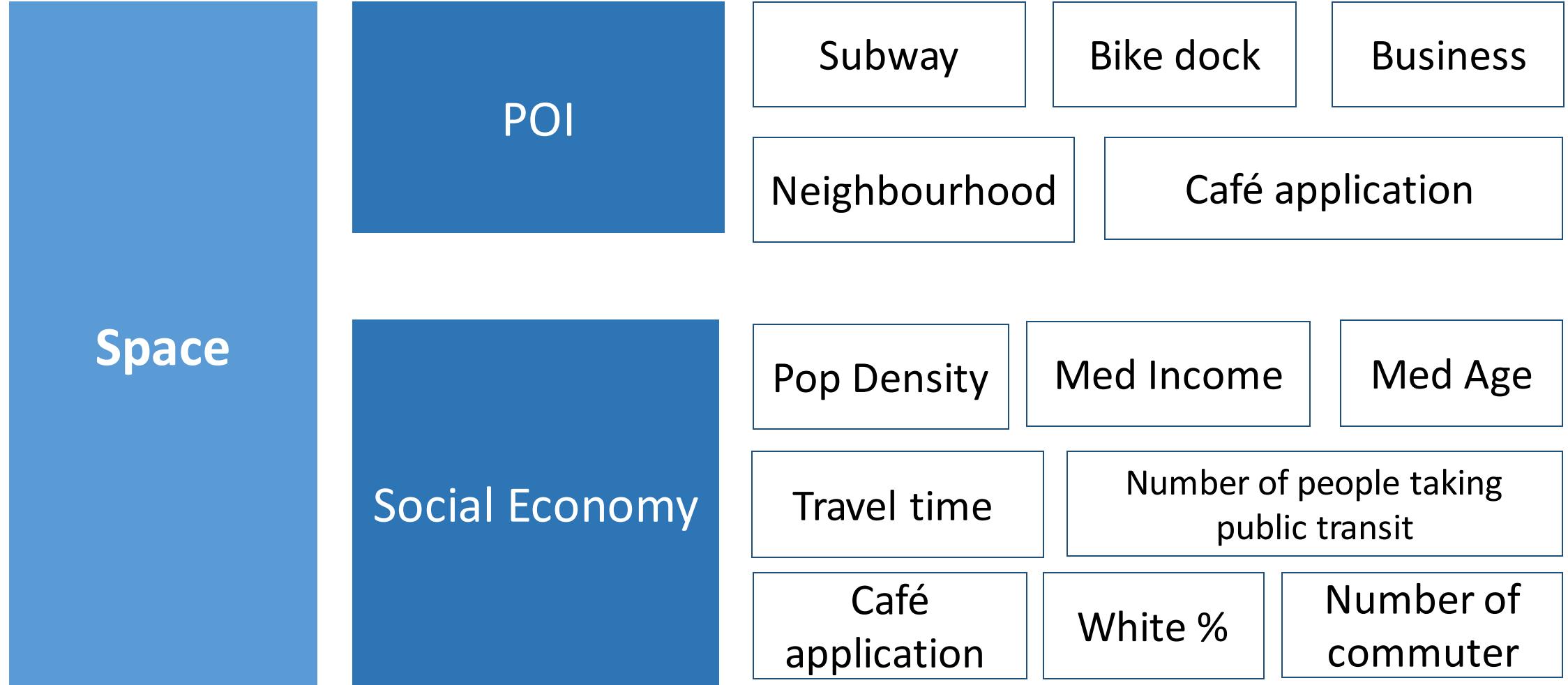
Day of Week

Time

NYC Uber Demand Related Variables



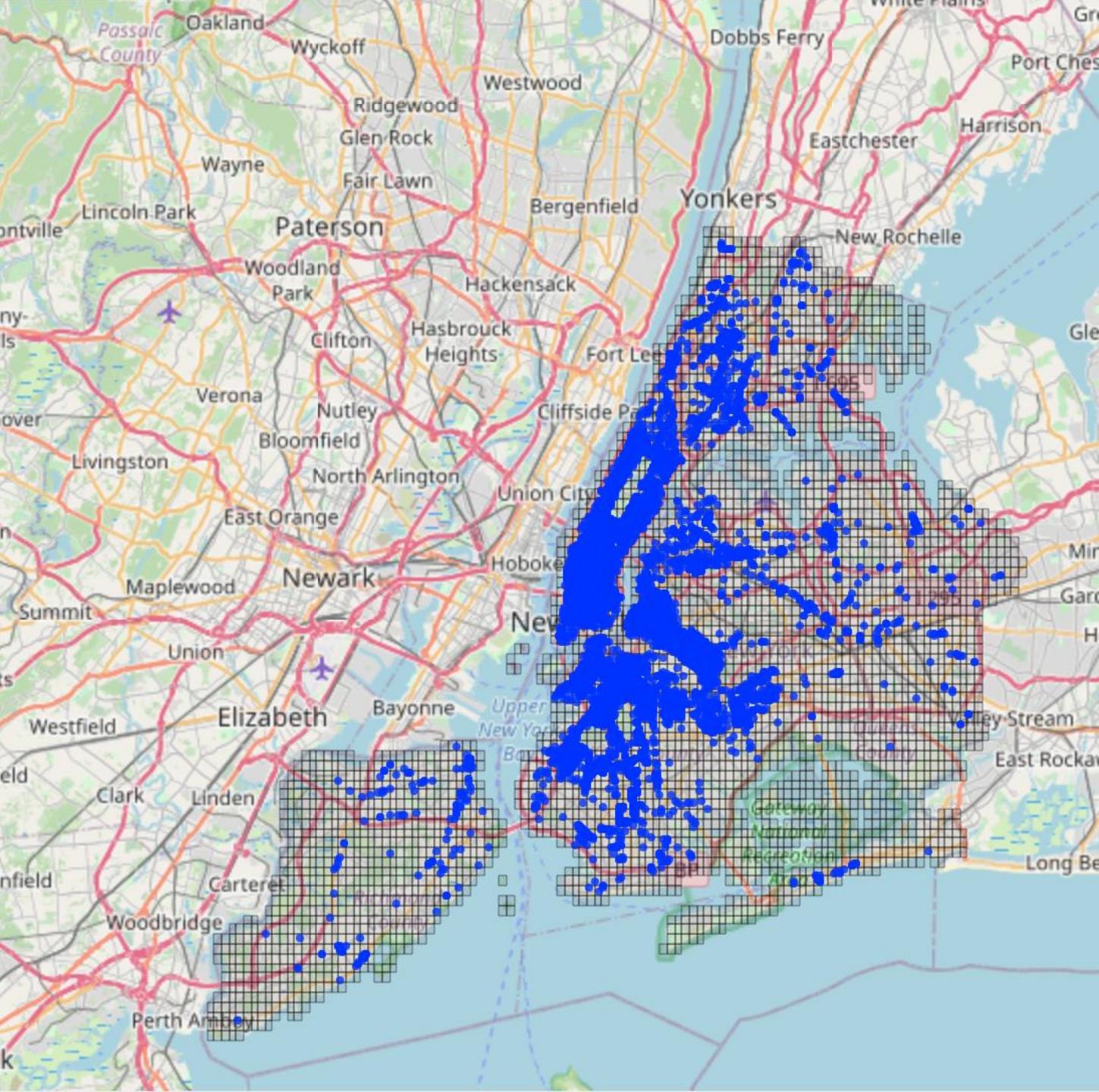
NYC Uber Demand Related Variables



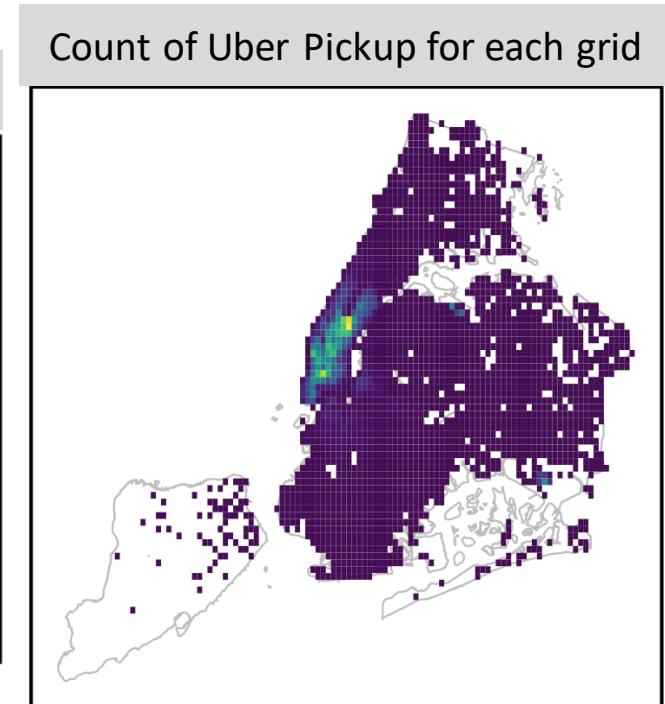
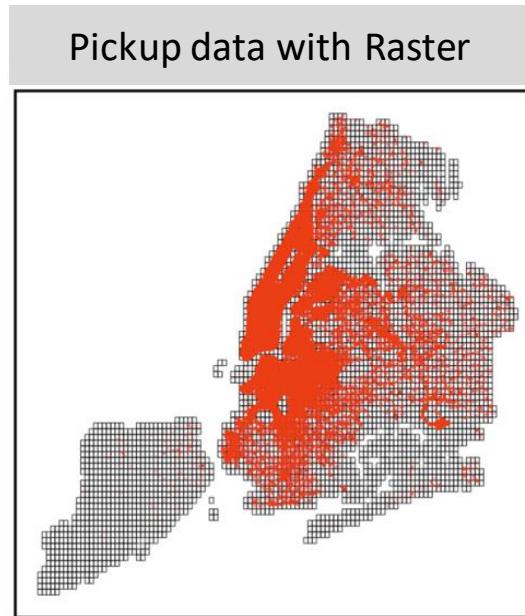
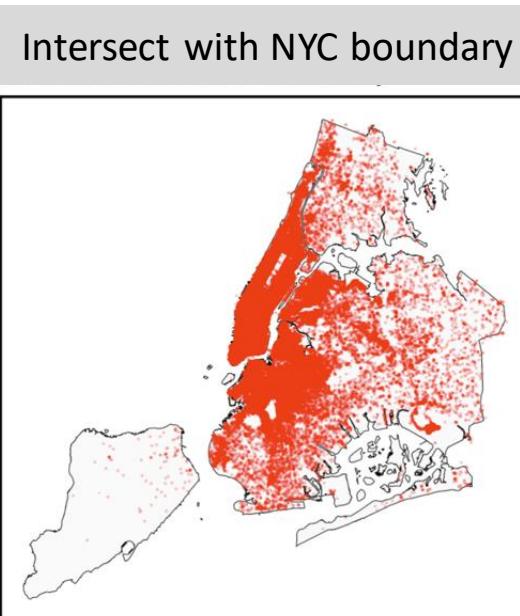
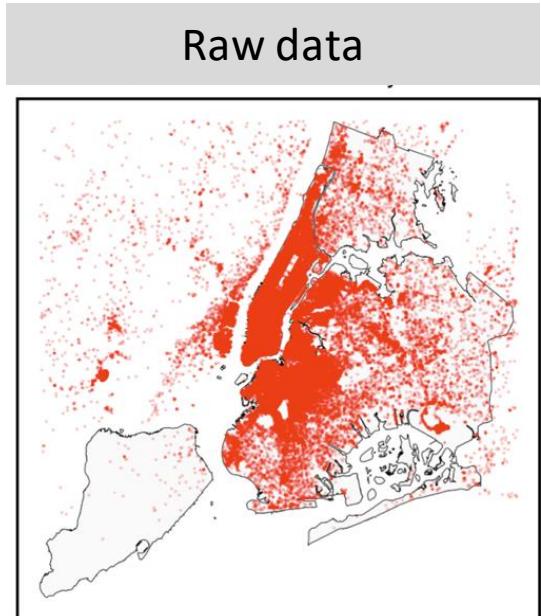
Related Variables Summary

| Variable | Data Source | Data Time | Data Size (row) |
|---------------------------|---------------|-----------|-----------------|
| Bike parking dock | NYC DOT | 2017 | 11734 |
| Subway Entries | NYC open data | 2018 | 1928 |
| Sidewalk Cafe application | NYC open data | 2017 | 1448 |
| Legal operating business | NYC open data | 2014 | 84,383 |
| American Community Survey | ACS | 2014 | - |
| Weather | | 2014-May | - |
| NYC boundary data | NYC open data | 2018 | 195 |

3. Exploratory Analysis



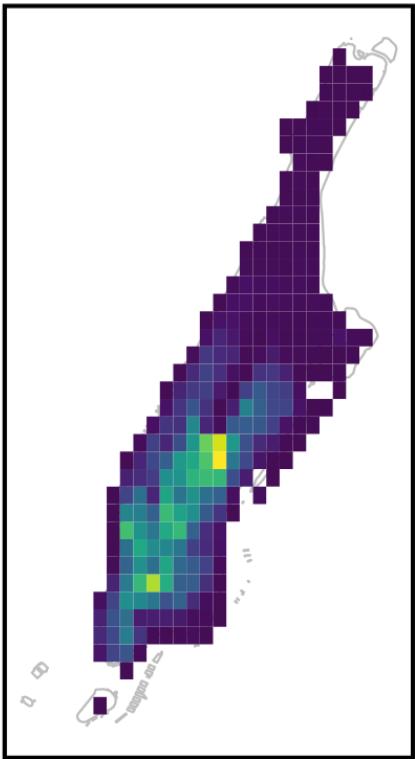
NYC Uber Pickup data



500*500m grid
In total 3949 grids

Zoom in to Manhattan

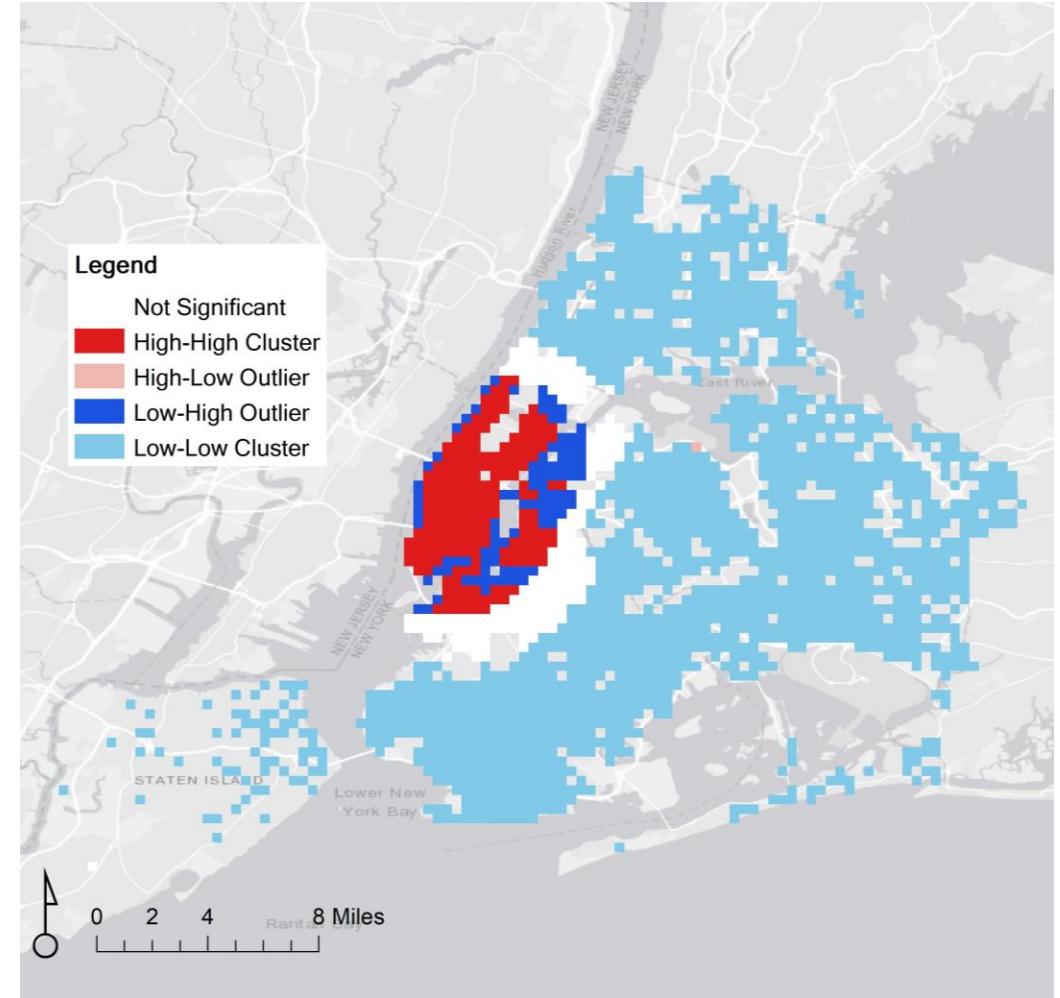
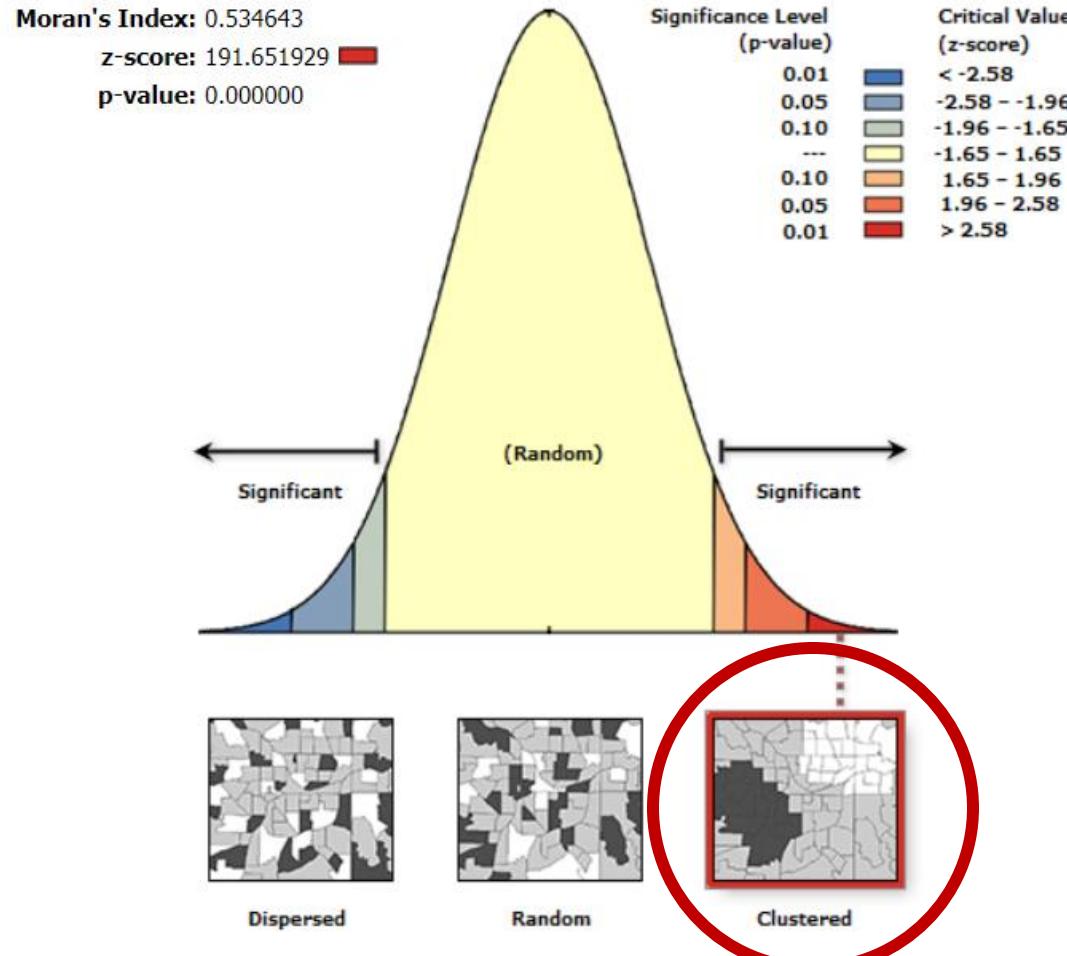
Manhattan Uber Demand



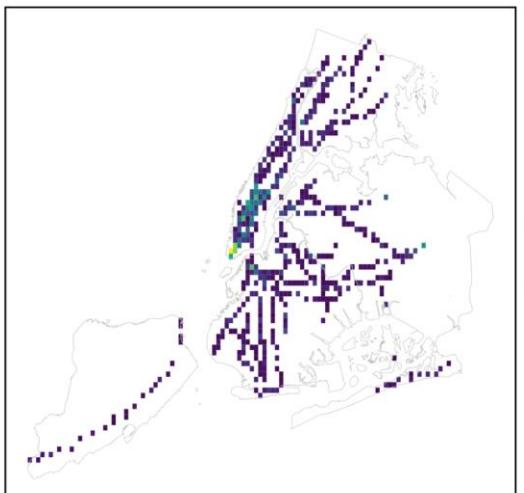
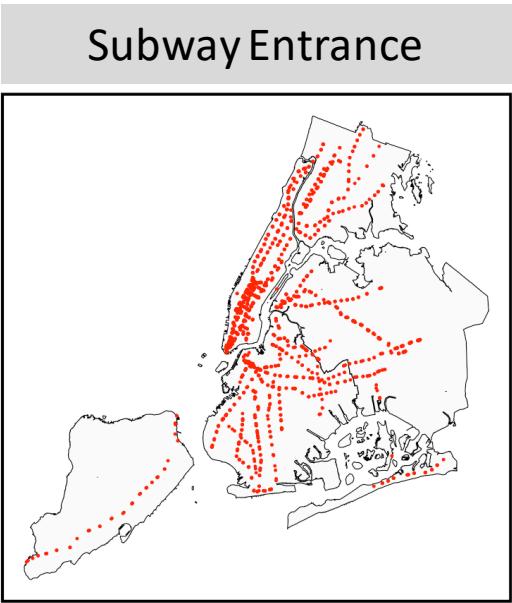
Top 10 Busy Uber Pickup Grid



NYC Uber Pickup - Spatial Cluster Analysis (Moran's I)

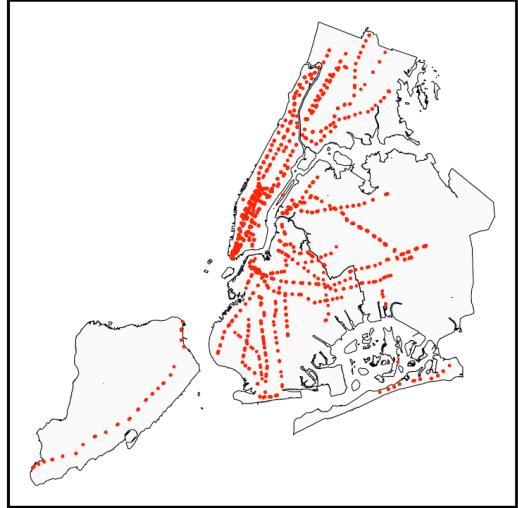


Exploratory Analysis – Location

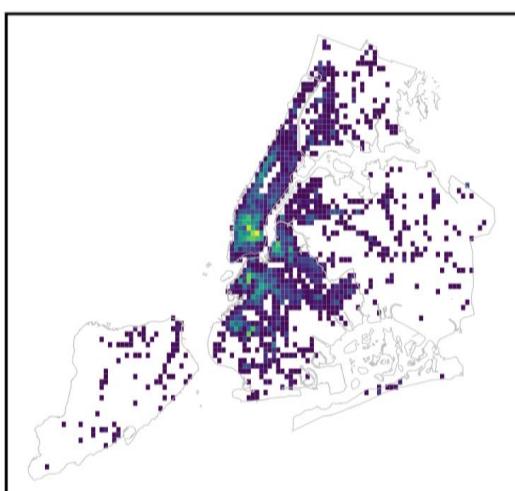
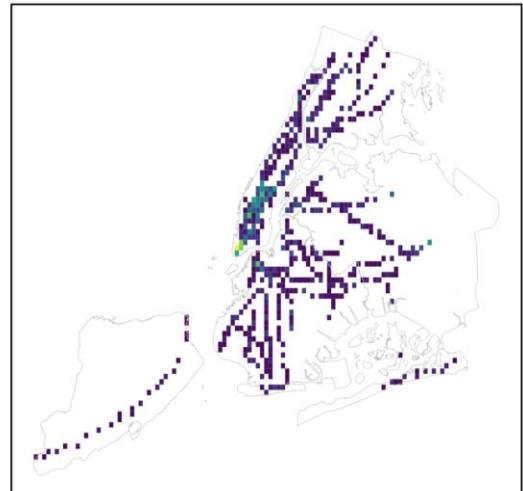
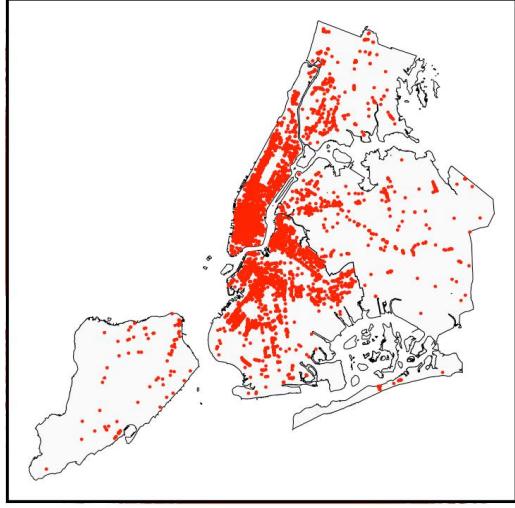


Exploratory Analysis – Location

Subway Entrance



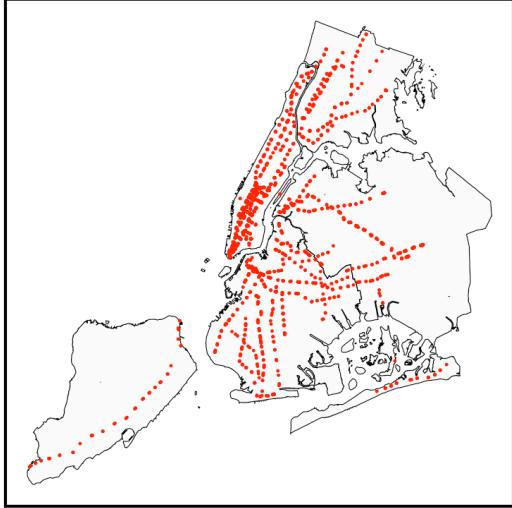
Bike Dock



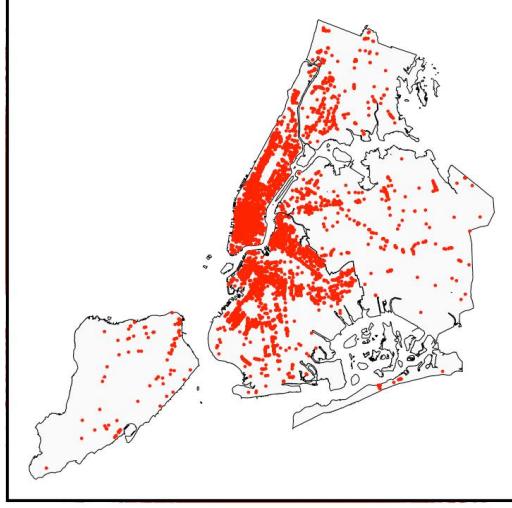
countBike
100
75
50
25

Exploratory Analysis – Location

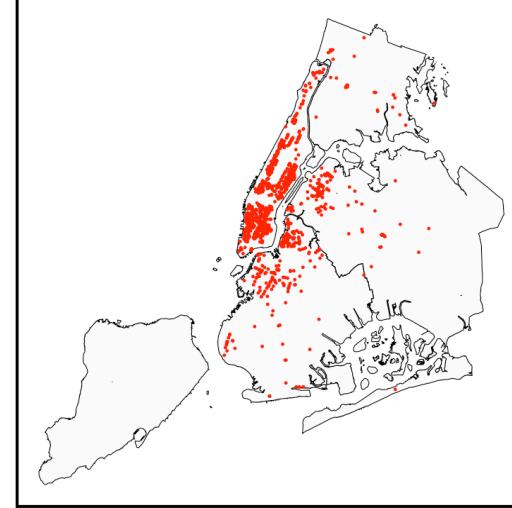
Subway Entrance



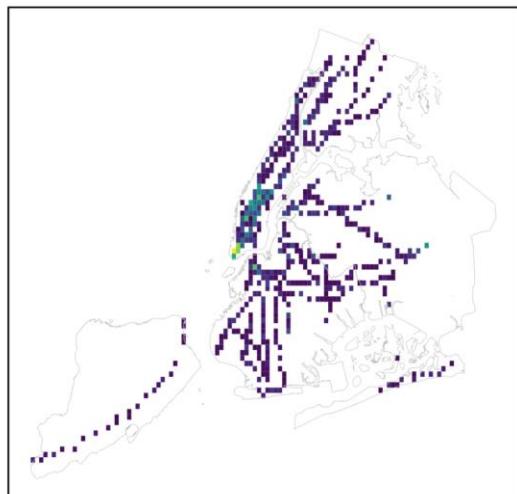
Bike Dock



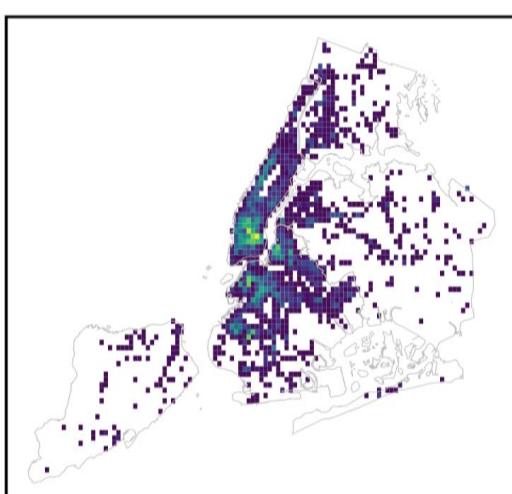
Café Application



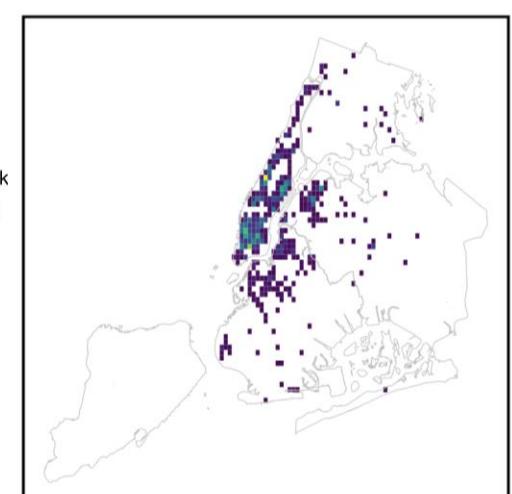
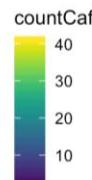
countSu



countBik

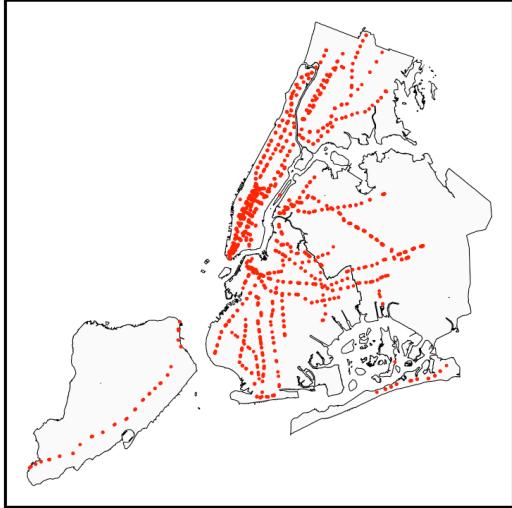


countCafe

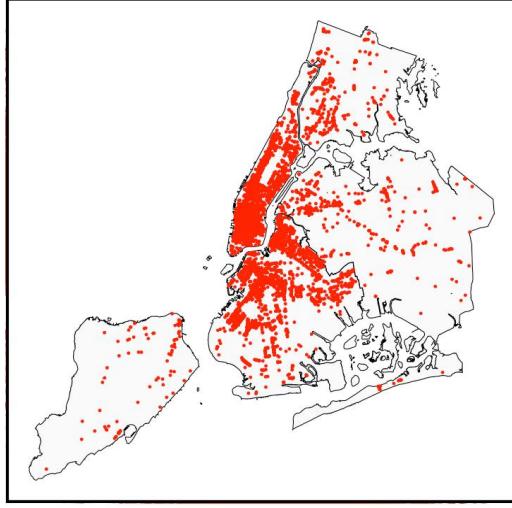


Exploratory Analysis – Location

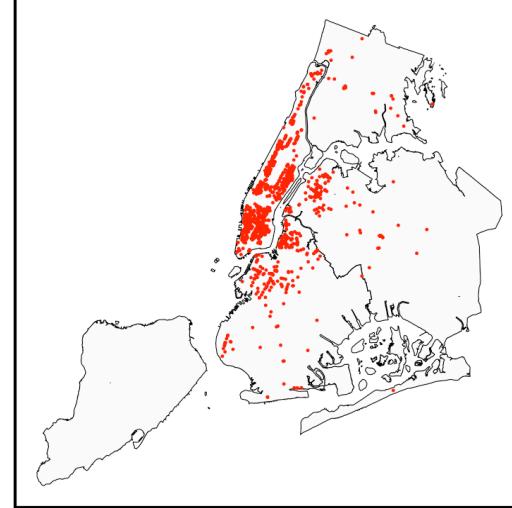
Subway Entrance



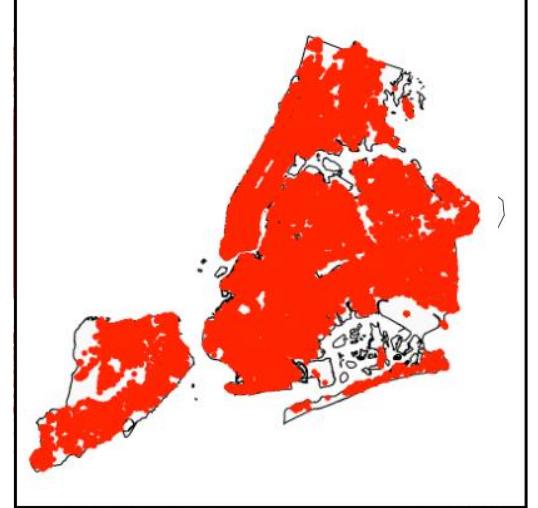
Bike Dock



Café Application

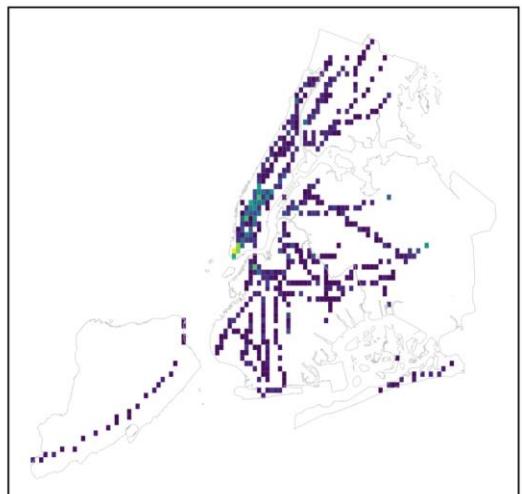


Legal Business



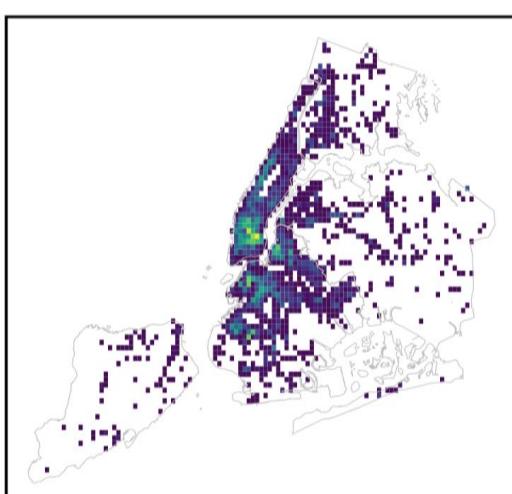
countSu

30
20
10



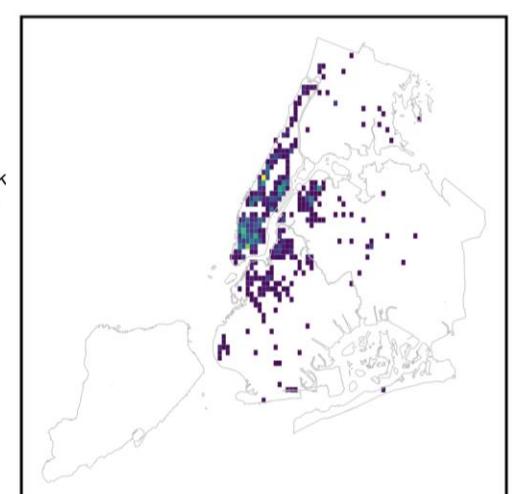
countBik

100
75
50
25



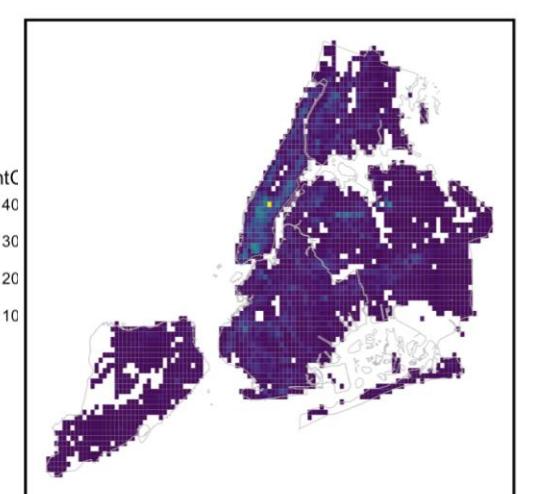
countC

40
30
20
10

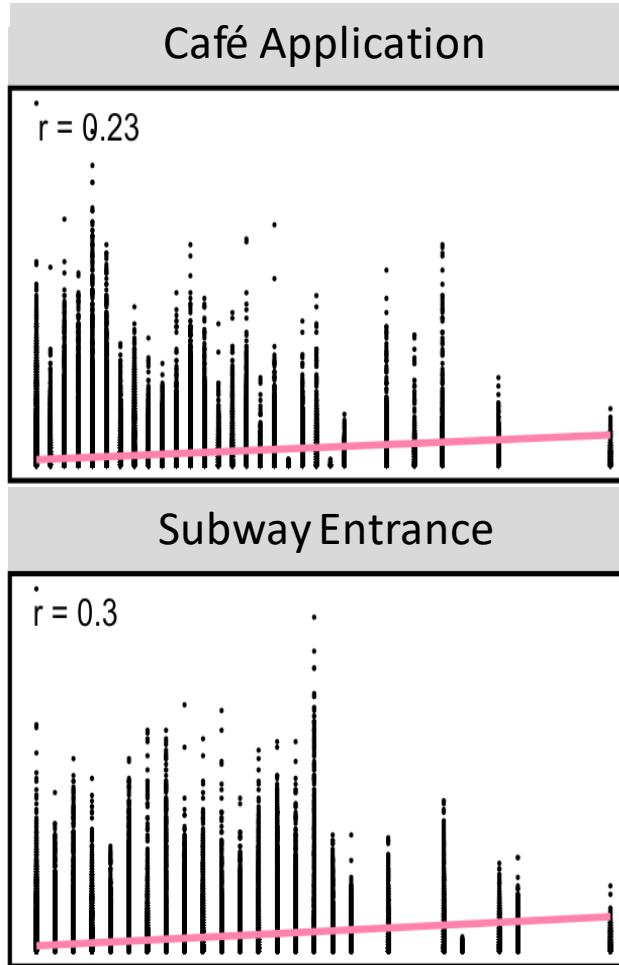
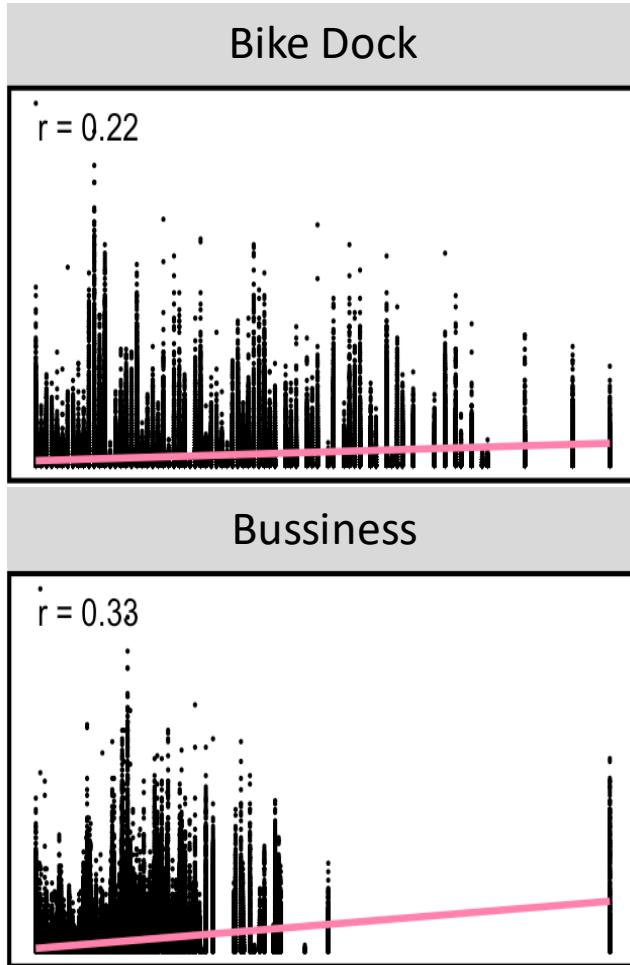


countPOI

600
400
200

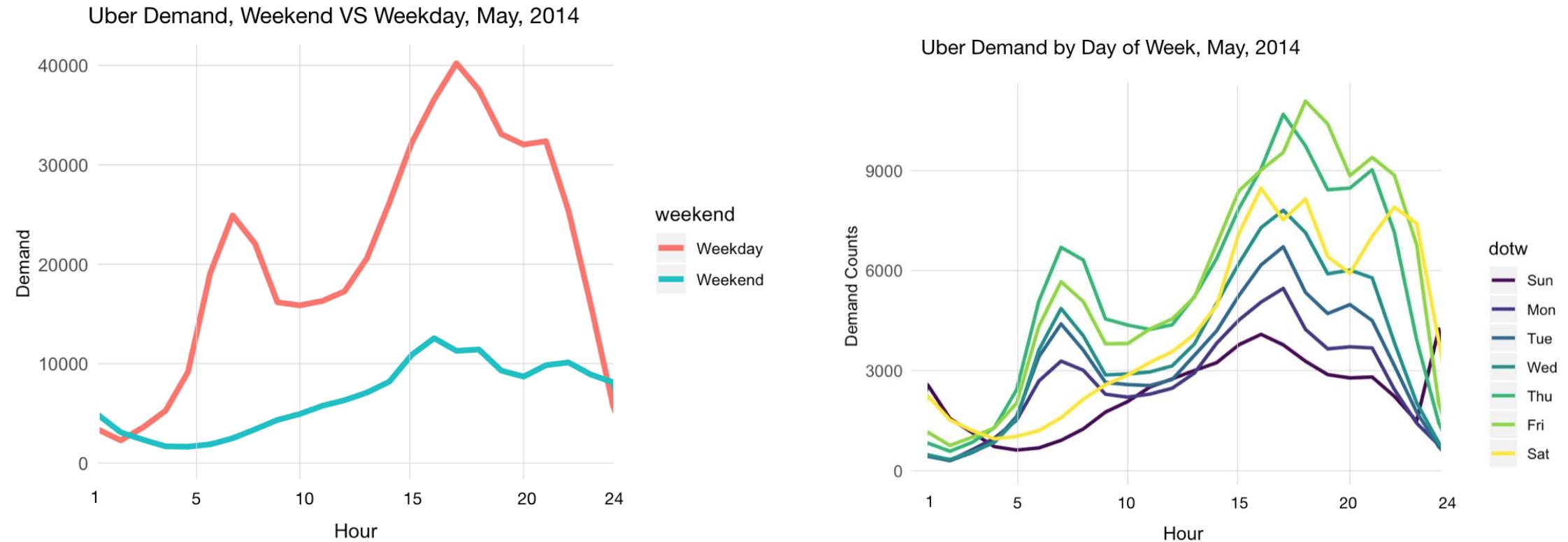


Exploratory Analysis – POI



The Demand for Uber has a **positive correlation** with the number of bike dock, subway entrance, business and café within the raster.

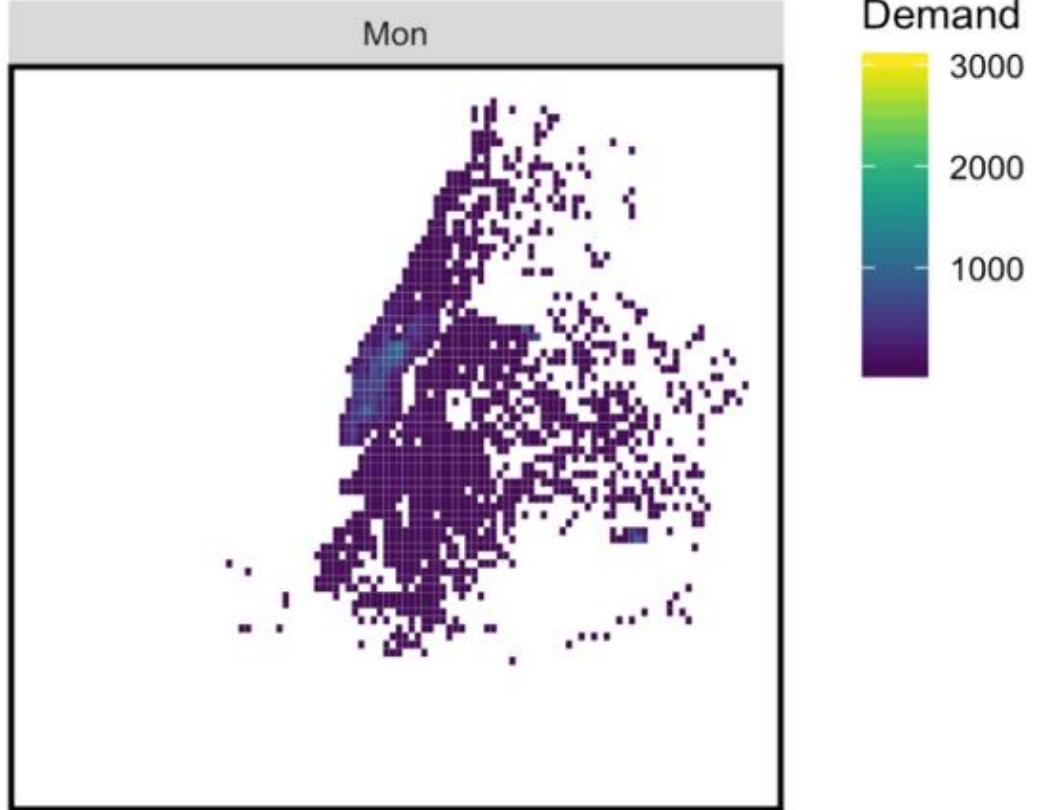
Exploratory Analysis – Time



The demand for Uber on weekdays is significantly higher than on weekends and shows a significant **AM and PM peak**, the demand is related to **commuting**.

From Monday to Friday, the demand for Uber gradually increases, and then decreases again at the weekend.

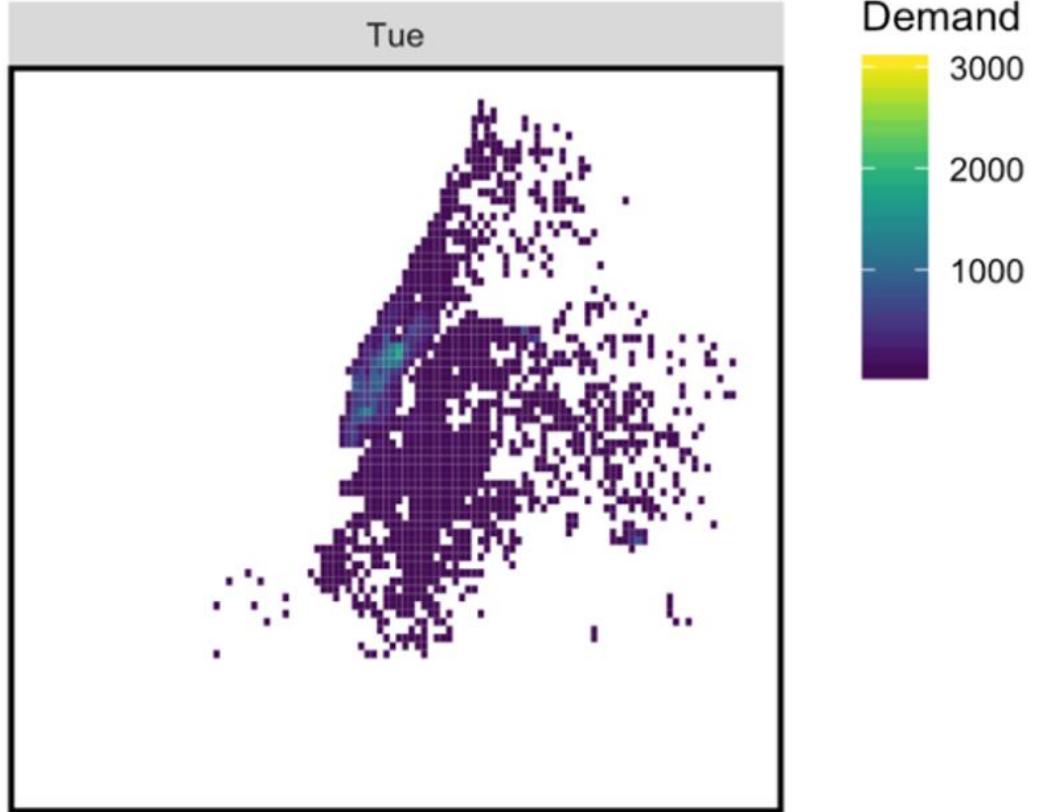
Exploratory Analysis – Time



Uber demand distribution evolved for different day of week.

For **weekday**, the demand hotspots remain the same, but for **Saturday** the hotspot changes, and for **Sunday** the demand is relative flat compared with other days.

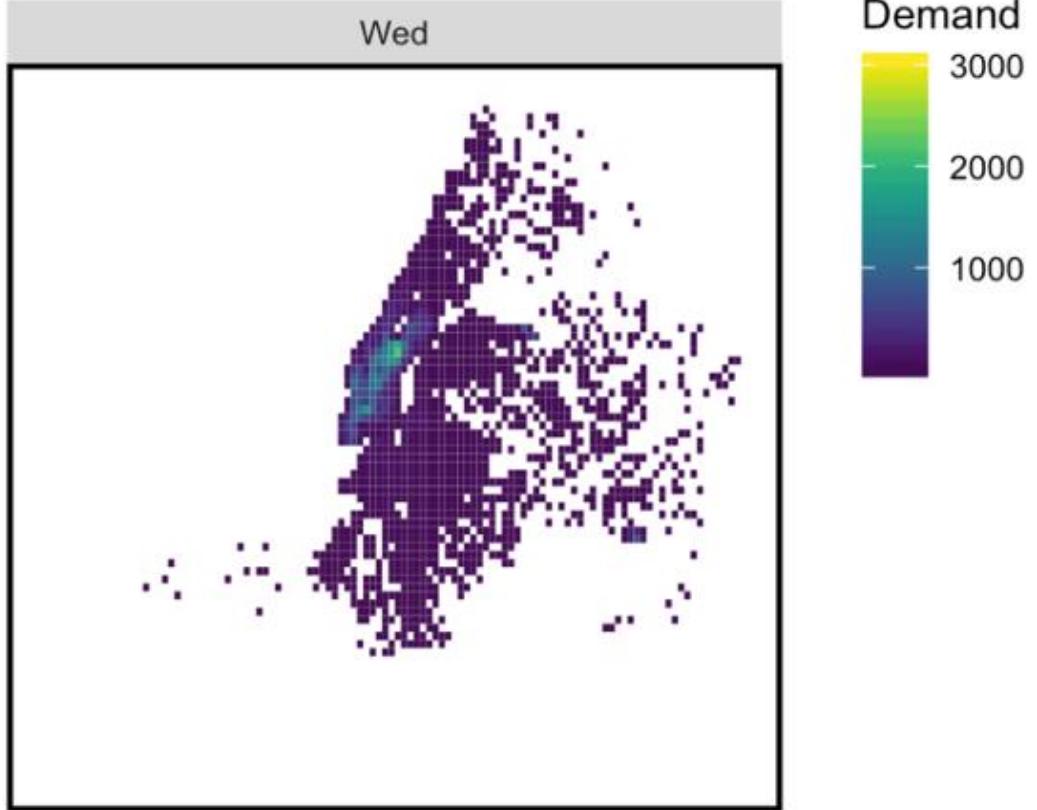
Exploratory Analysis – Time



Uber demand distribution evolved for different day of week.

For **weekday**, the demand hotspots remain the same, but for **Saturday** the hotspot changes, and for **Sunday** the demand is relative flat compared with other days.

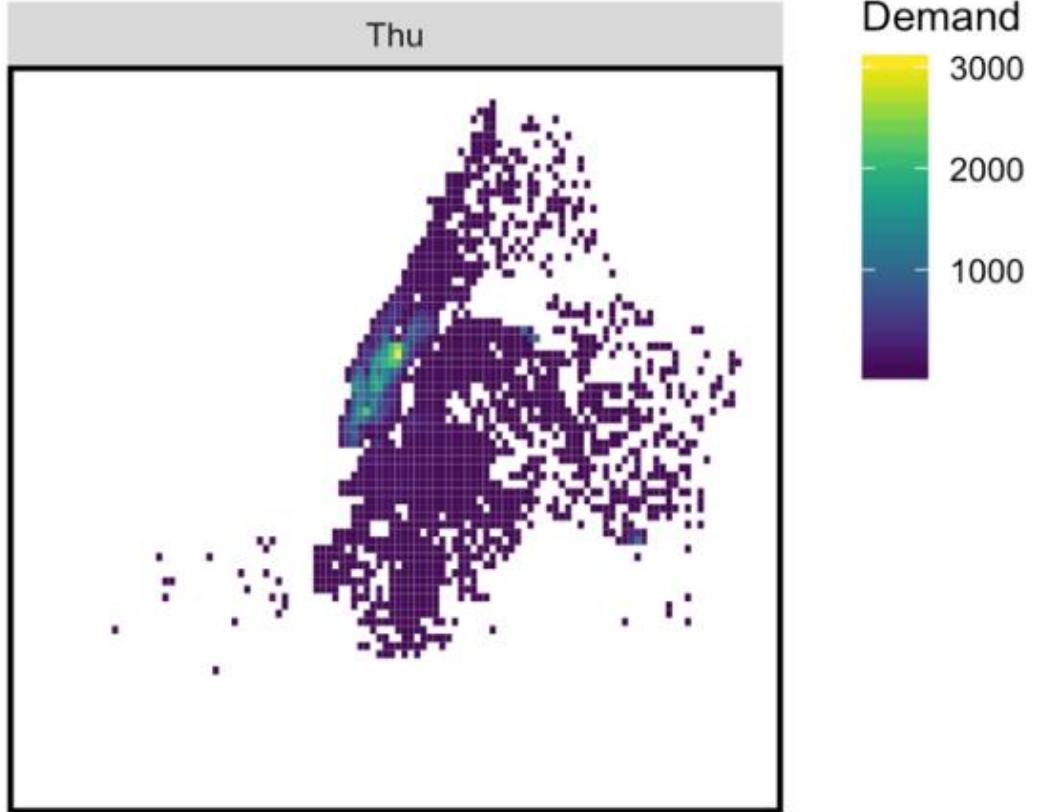
Exploratory Analysis – Time



Uber demand distribution evolved for different day of week.

For **weekday**, the demand hotspots remain the same, but for **Saturday** the hotspot changes, and for **Sunday** the demand is relative flat compared with other days.

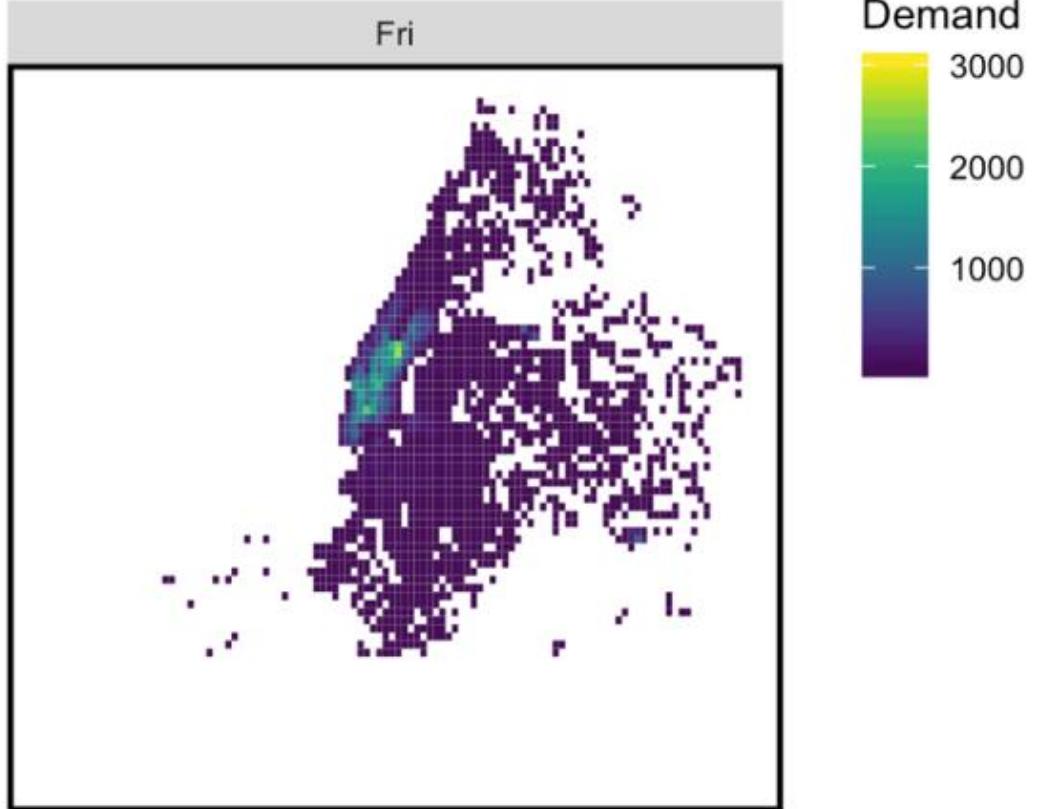
Exploratory Analysis – Time



Uber demand distribution evolved for different day of week.

For **weekday**, the demand hotspots remain the same, but for **Saturday** the hotspot changes, and for **Sunday** the demand is relative flat compared with other days.

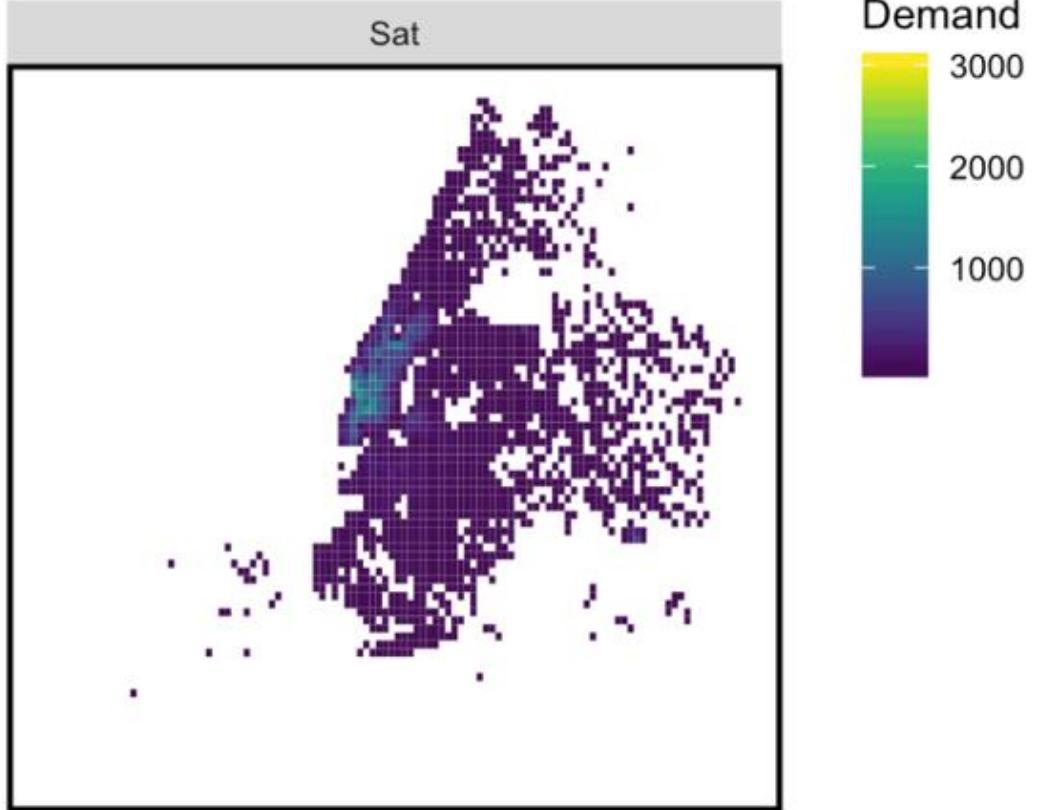
Exploratory Analysis – Time



Uber demand distribution evolved for different day of week.

For **weekday**, the demand hotspots remain the same, but for **Saturday** the hotspot changes, and for **Sunday** the demand is relative flat compared with other days.

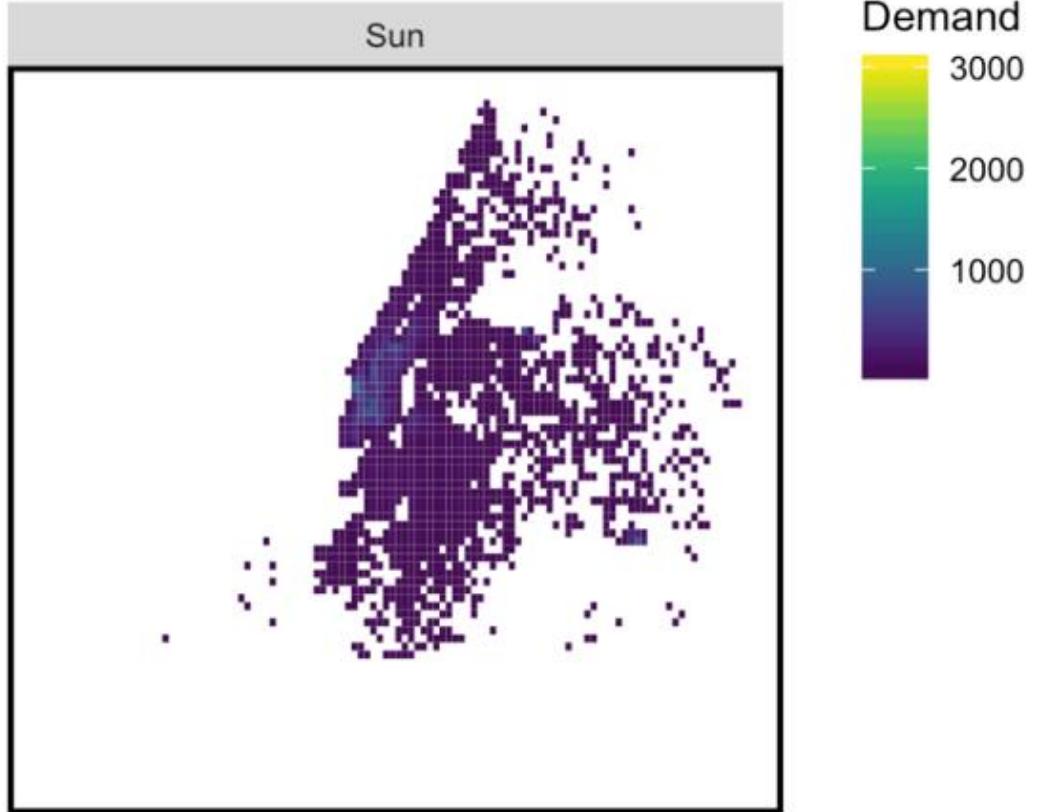
Exploratory Analysis – Time



Uber demand distribution evolved for different day of week.

For **weekday**, the demand hotspots remain the same, but for **Saturday** the hotspot changes, and for **Sunday** the demand is relative flat compared with other days.

Exploratory Analysis – Time



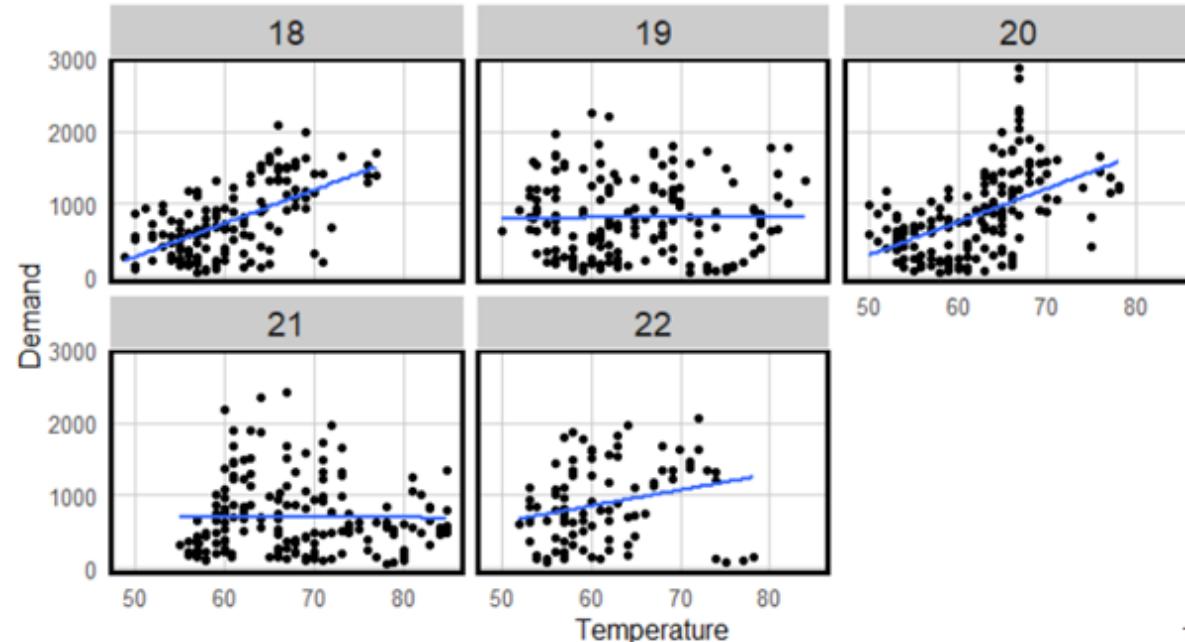
Uber demand distribution evolved for different day of week.

For **weekday**, the demand hotspots remain the same, but for **Saturday** the hotspot changes, and for **Sunday** the demand is relative flat compared with other days.

Exploratory Analysis – Weather

Demand as a function of Temperature by week

Demand by week; May, 2014

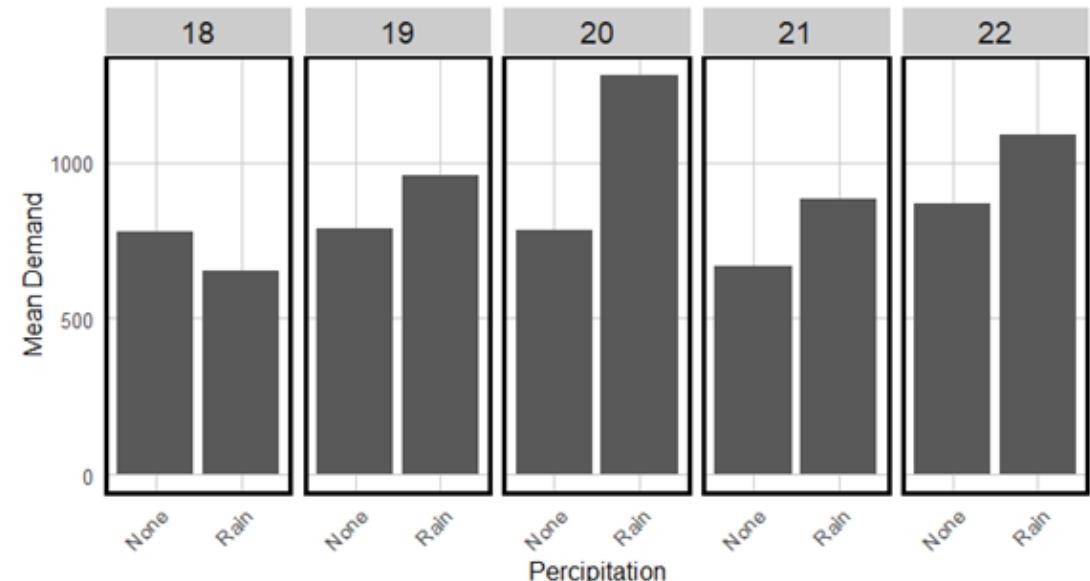


The demand for Uber increases slightly as **temperature warm**

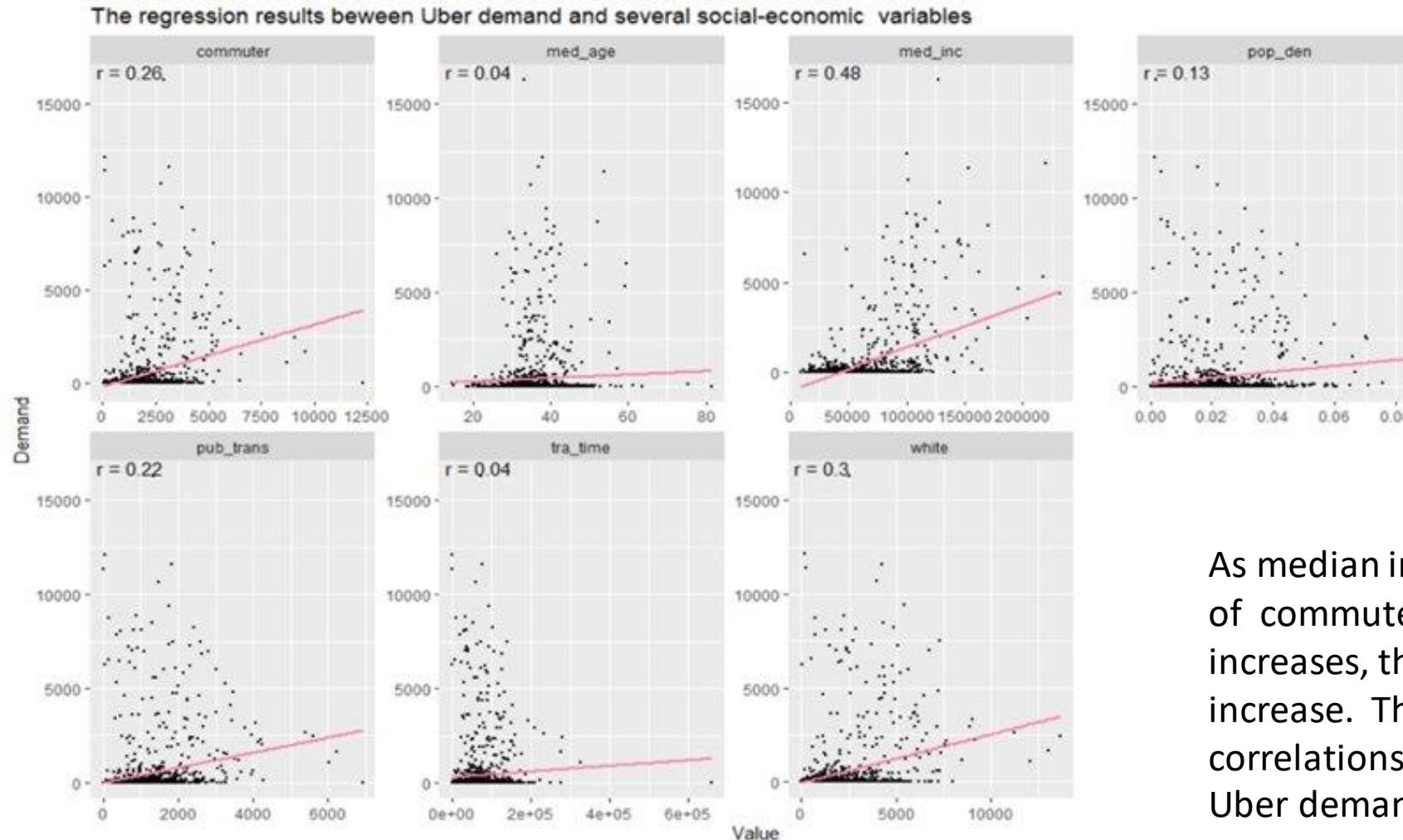
The demand for Uber is higher on raining days, **precipitation** increase the propensity to take Uber

Does Uber demand vary when it's raining?

Mean Demand by week; May, 2014

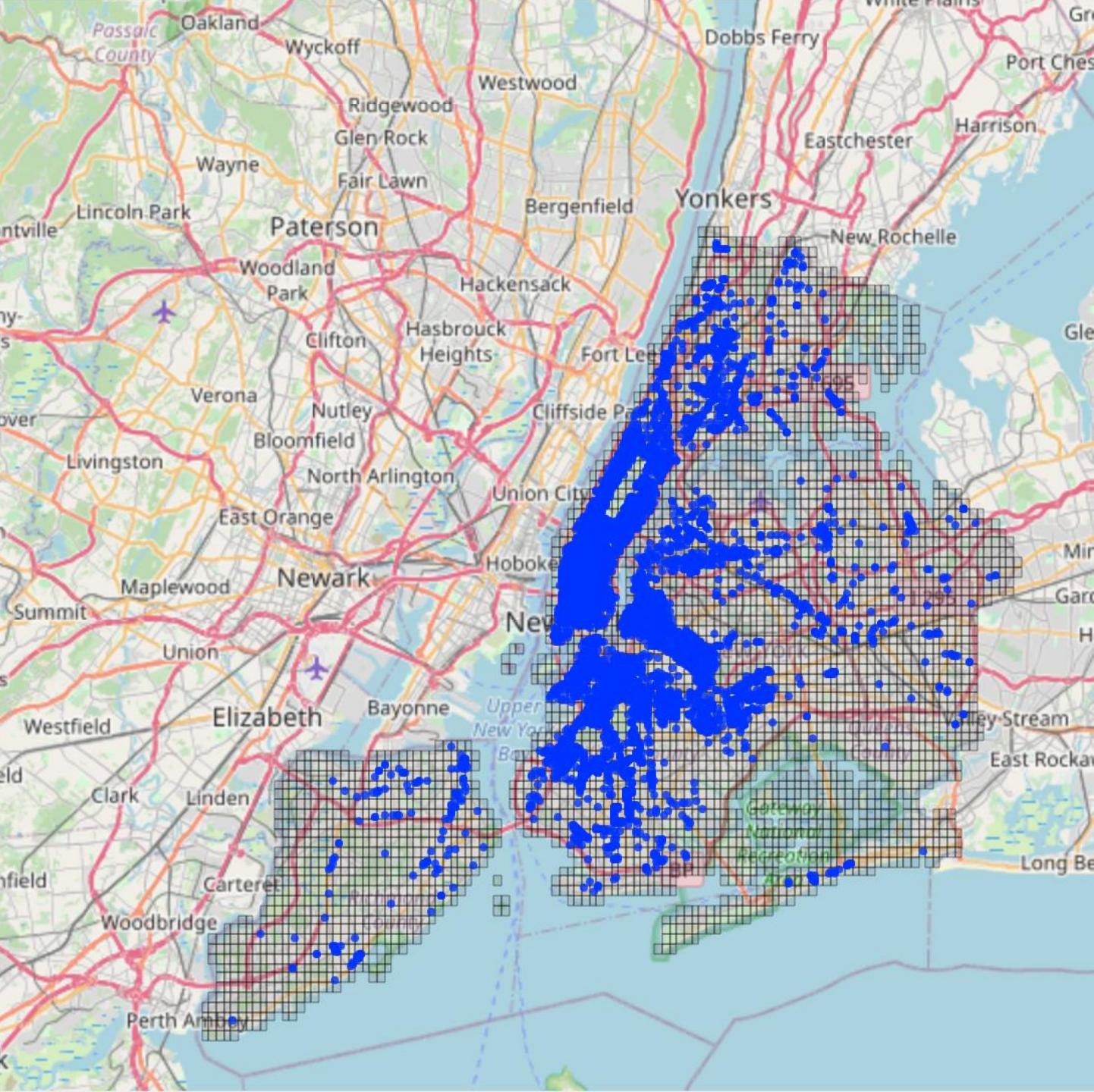


Exploratory Analysis – Social Economic Characteristic



As median income, the number of commuter, and percent of white increases, the demand for Uber increase. There is no significant correlation relationship between travel time and Uber demand

4. Model



Model Result – Forward and Backward Stepping

Call:

```
lm(formula = Demand ~ countPOI + hour + med_inc + dotw + tra_time +  
  commuter + countBike + countSubway + pub_trans + med_age +  
  countCafe + Temperature + pop_den + Wind_Speed + Percipitation +  
  white, data = uber)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---------|--------|--------|-------|---------|
| -13.106 | -2.475 | -0.714 | 1.221 | 125.338 |

No Variable be dropped by "Forward & Backward Stepping", which means that all the variables that we put are useful to explain the Uber demand.

Model Result – Summary Table

```
Call:  
lm(formula = Demand ~ pop + med_age + med_inc + white + commuter +  
    tra_time + pub_trans + pop_den + Temperature + isPercip +  
    Wind_Speed + countBike + countSubway + countCafe + countPOI +  
    dotw + Day_hour + boro_name + ntaname, data = all2)
```

Residuals:

| Min | 1Q | Median | 3Q | Max |
|---------|--------|--------|-------|---------|
| -13.609 | -2.198 | -0.490 | 1.279 | 124.667 |

Coefficients: (1 not defined because of singularities)

| | Estimate | Std. Error | t value | Pr(> t) |
|-------------|------------|------------|---------|--------------|
| (Intercept) | 4.224e+00 | 6.137e-01 | 6.883 | 5.89e-12 *** |
| pop | -1.656e-04 | 1.808e-05 | -9.162 | < 2e-16 *** |
| med_age | 4.606e-02 | 2.114e-03 | 21.792 | < 2e-16 *** |
| med_inc | 1.136e-05 | 4.728e-07 | 24.022 | < 2e-16 *** |
| white | 6.227e-05 | 2.007e-05 | 3.103 | 0.001919 ** |
| commuter | -3.093e-04 | 5.665e-05 | -5.460 | 4.76e-08 *** |
| tra_time | -2.571e-06 | 7.661e-07 | -3.356 | 0.000791 *** |
| pub_trans | 4.514e-04 | 5.816e-05 | 7.761 | 8.46e-15 *** |
| pop_den | 7.962e+00 | 1.494e+00 | 5.327 | 9.97e-08 *** |
| Temperature | -3.076e-02 | 2.066e-03 | -14.890 | < 2e-16 *** |
| isPercip1 | 5.135e-01 | 3.801e-02 | 13.510 | < 2e-16 *** |
| Wind_Speed | -2.420e-02 | 4.179e-03 | -5.791 | 7.02e-09 *** |
| countBike | 1.348e-02 | 8.710e-04 | 15.480 | < 2e-16 *** |
| countSubway | 7.712e-02 | 3.083e-03 | 25.016 | < 2e-16 *** |
| countCafe | 5.235e-02 | 2.751e-03 | 19.027 | < 2e-16 *** |
| countPOI | 5.042e-03 | 2.614e-04 | 19.291 | < 2e-16 *** |
| dotwMon | -1.241e+00 | 4.619e-02 | -26.860 | < 2e-16 *** |
| dotwSat | -9.888e-01 | 3.966e-02 | -24.929 | < 2e-16 *** |
| dotwSun | -1.769e+00 | 4.397e-02 | -40.230 | < 2e-16 *** |
| dotwThu | -3.921e-02 | 4.118e-02 | -0.952 | 0.341017 |
| dotwTue | -6.581e-01 | 4.632e-02 | -14.206 | < 2e-16 *** |
| dotwWed | -3.168e-01 | 4.453e-02 | -7.115 | 1.13e-12 *** |
| Day_hour1 | -8.051e-01 | 1.066e-01 | -7.556 | 4.19e-14 *** |
| Day_hour10 | 2.194e-01 | 8.999e-02 | 2.438 | 0.014774 * |
| Day_hour11 | 4.367e-01 | 8.966e-02 | 4.870 | 1.11e-06 *** |

- All the variables are **significant** (the screenshot omit some variables like Day_hour and neighbourhood)
- While the **R^2 is relatively low**, more related variables need to be added to explain the Uber demand.

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

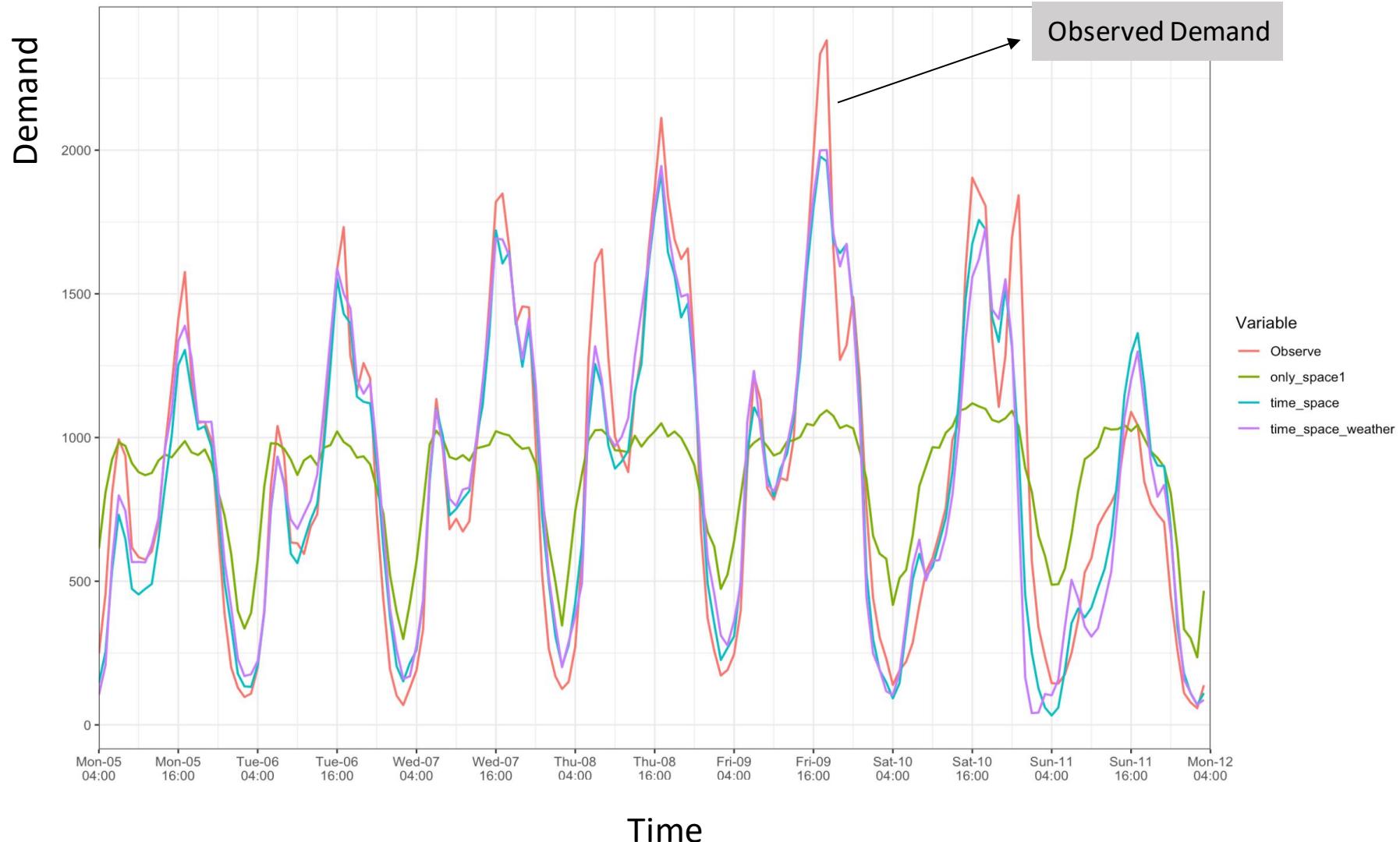
Residual standard error: 4.77 on 164228 degrees of freedom

Multiple R-squared: 0.3508, Adjusted R-squared: 0.3499

F-statistic: 380.9 on 233 and 164228 DF, p-value: < 2.2e-16

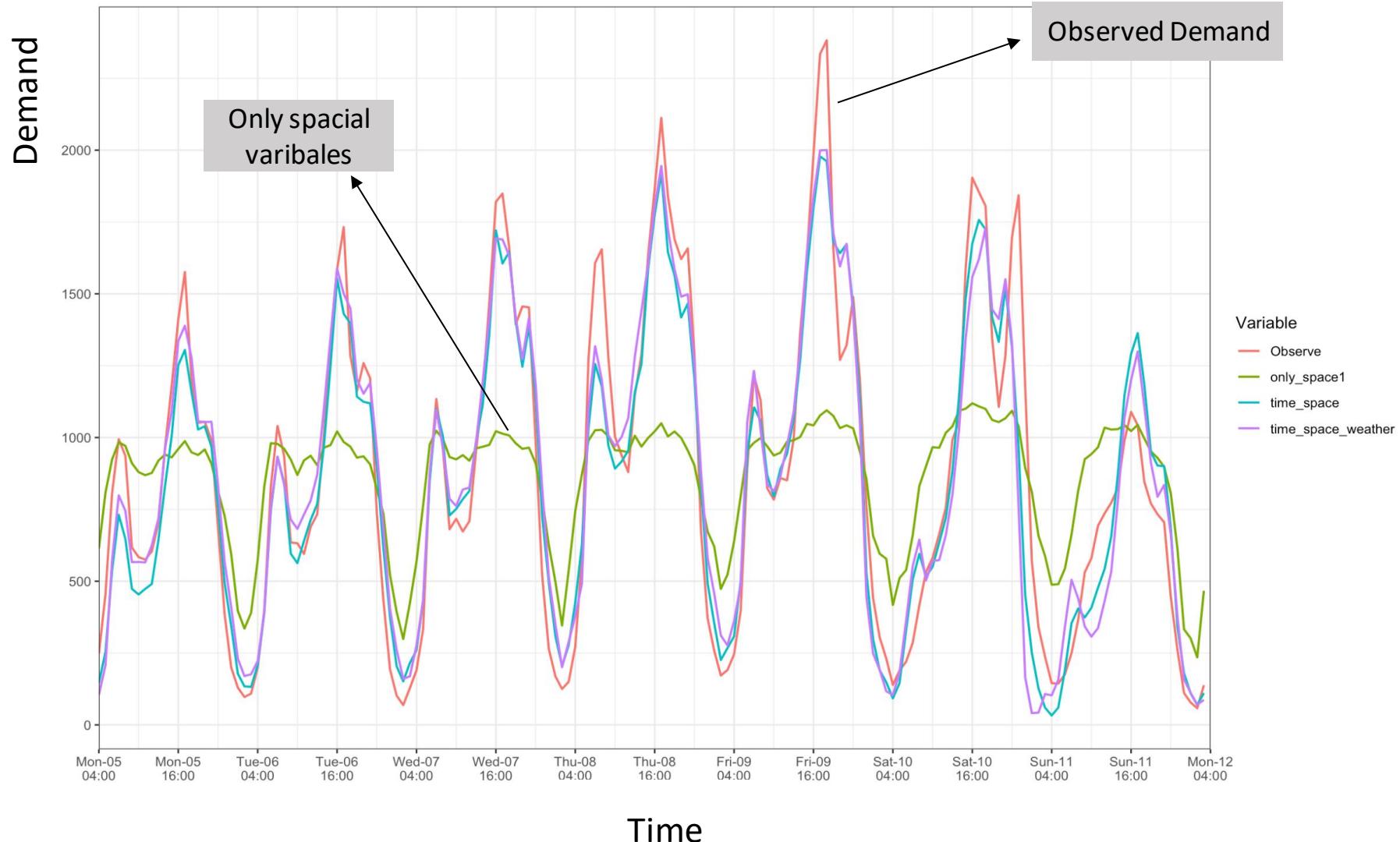
Model Result

Model Accuracy Comparison (only from 2014-05-07 to 2014-05-14)



Model Result

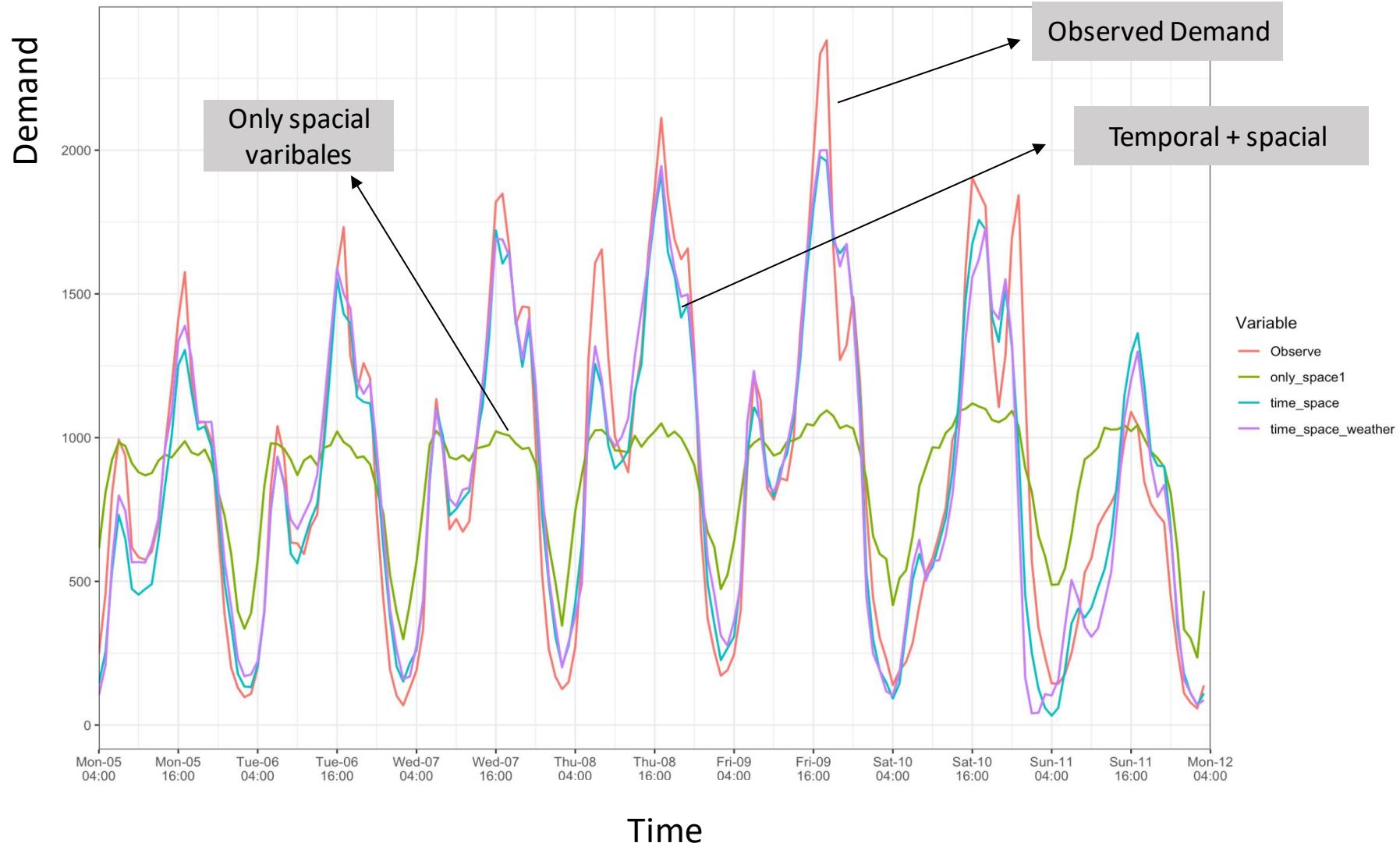
Model Accuracy Comparison (only from 2014-05-07 to 2014-05-14)



Spacial varibales (e.g. neighbourhood, POI, subway entries) can't explain all the changes of Uber demand.

Model Result

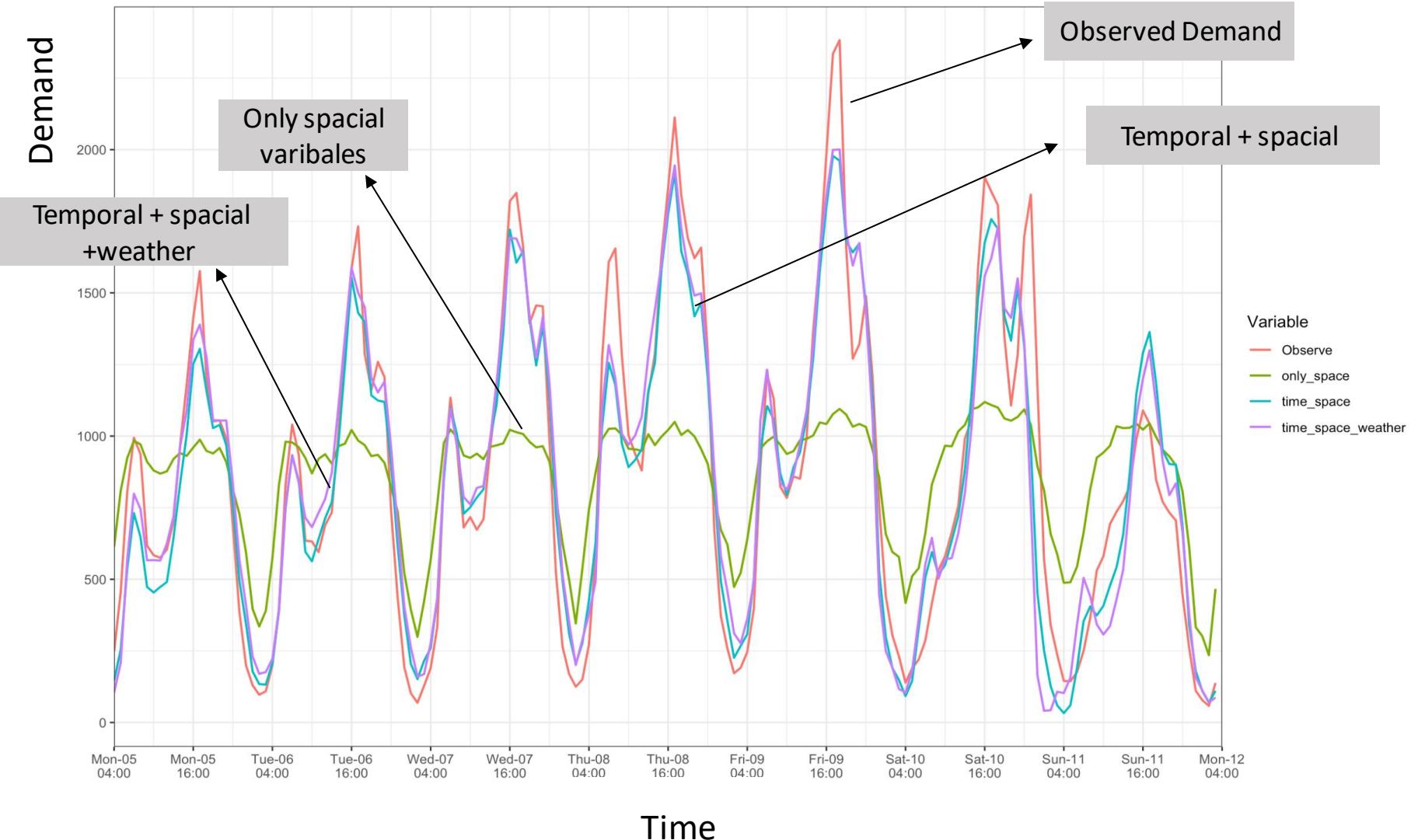
Model Accuracy Comparison (only from 2014-05-07 to 2014-05-14)



Additional Temporal variables
(e.g. day of week, hour of day)
can explain the most of fluctuant Uber demand

Model Result

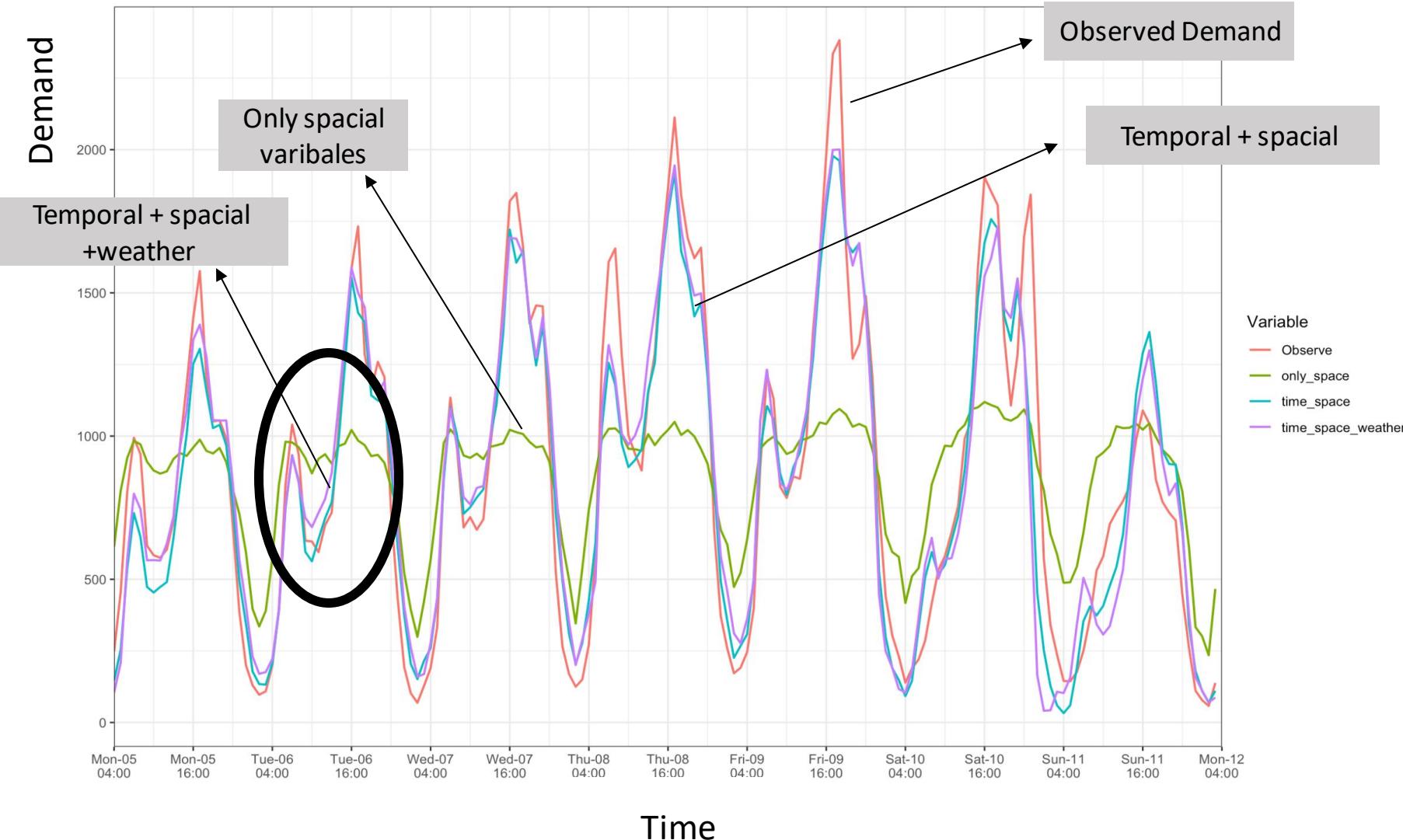
Model Accuracy Comparison (only from 2014-05-07 to 2014-05-14)



The predictive ability of **Weather related varibales** (e.g. Wind, precipitation) is unstable.

Model Result

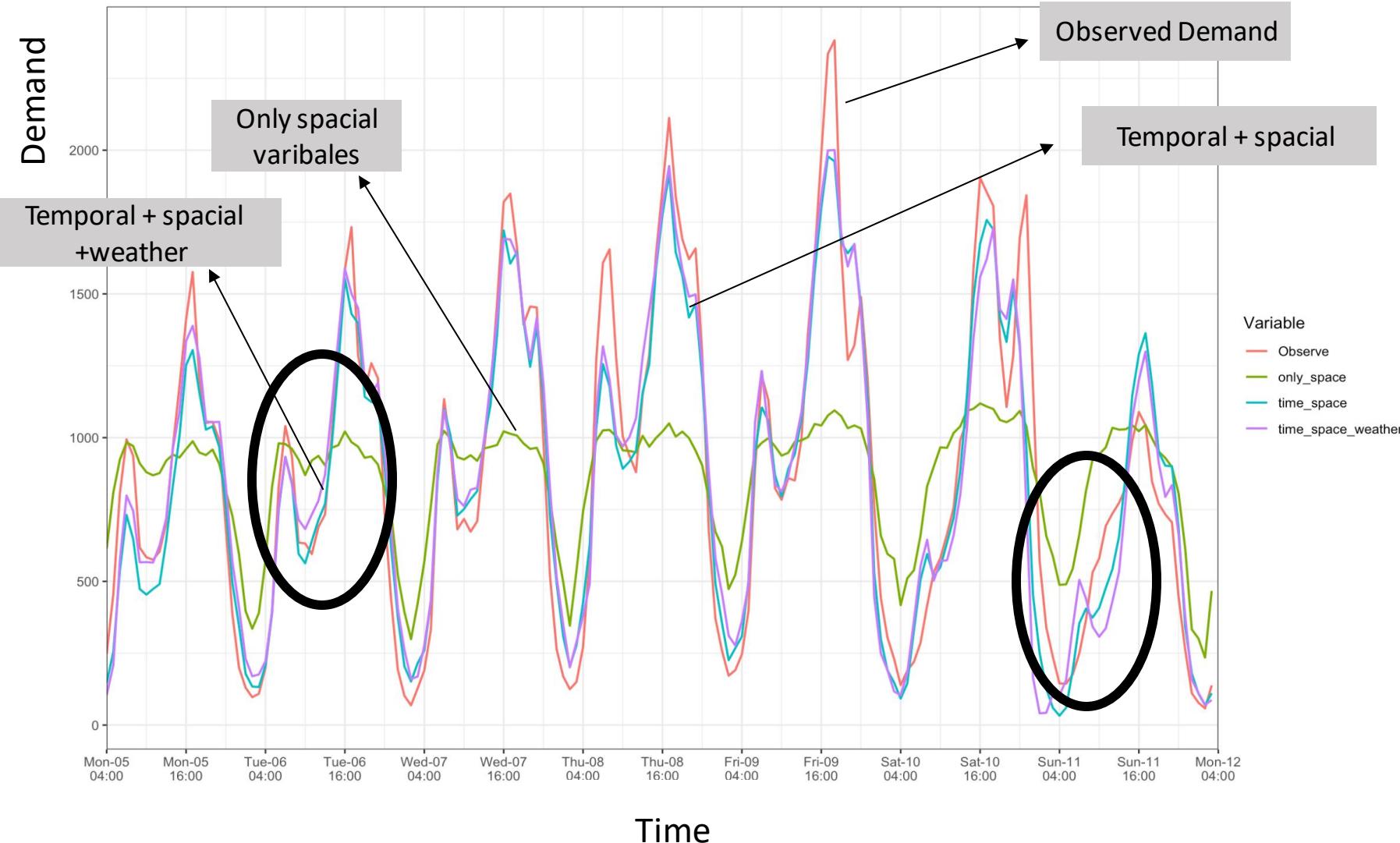
Model Accuracy Comparison (only from 2014-05-07 to 2014-05-14)



The predictive ability of **Weather related varibales** (e.g. Wind, precipitation) is unstable. It increase the accuracy in some days,

Model Result

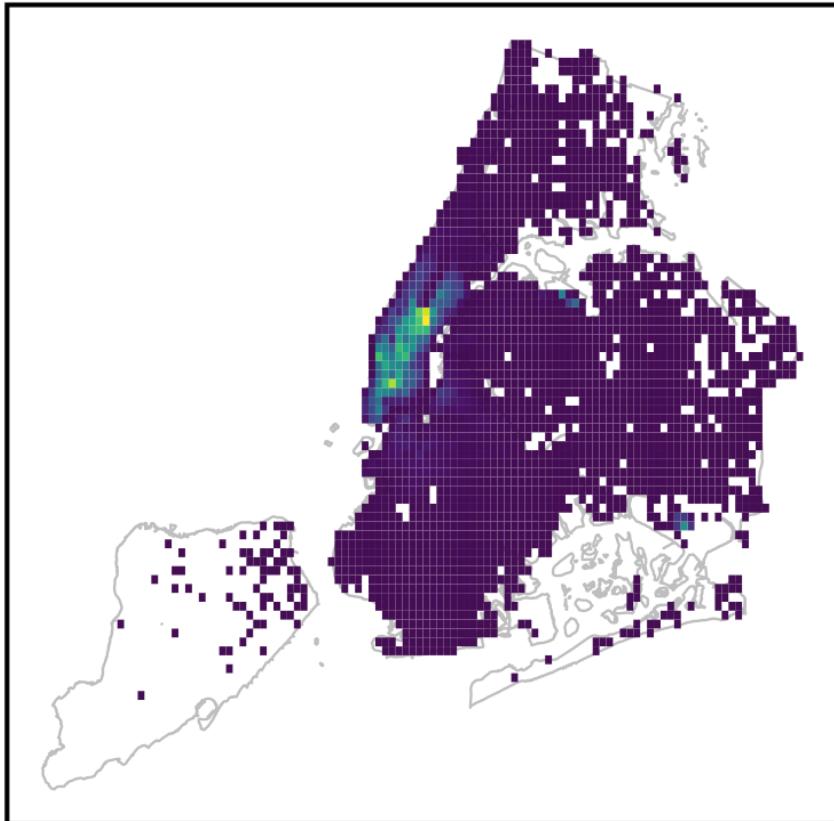
Model Accuracy Comparison (only from 2014-05-07 to 2014-05-14)



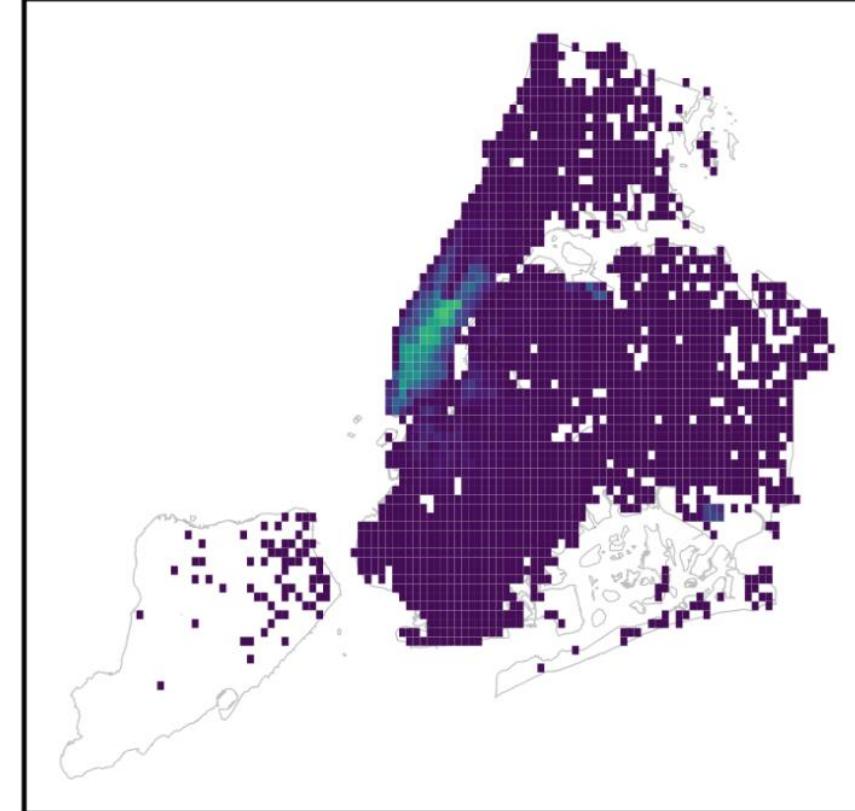
The predictive ability of **Weather related varibales** (e.g. Wind, precipitation) is unstable. It increase the accuracy in some days, while reduce the accuracy in some cases.

Model Result – Spacial Accuracy

Observe Uber Demand

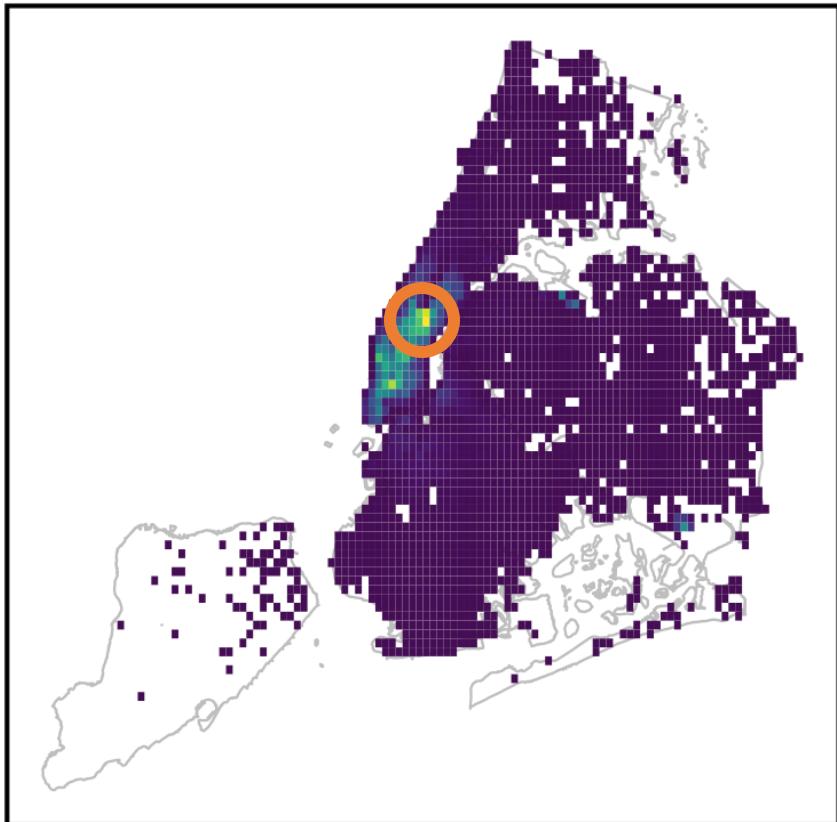


Predicted Uber Demand

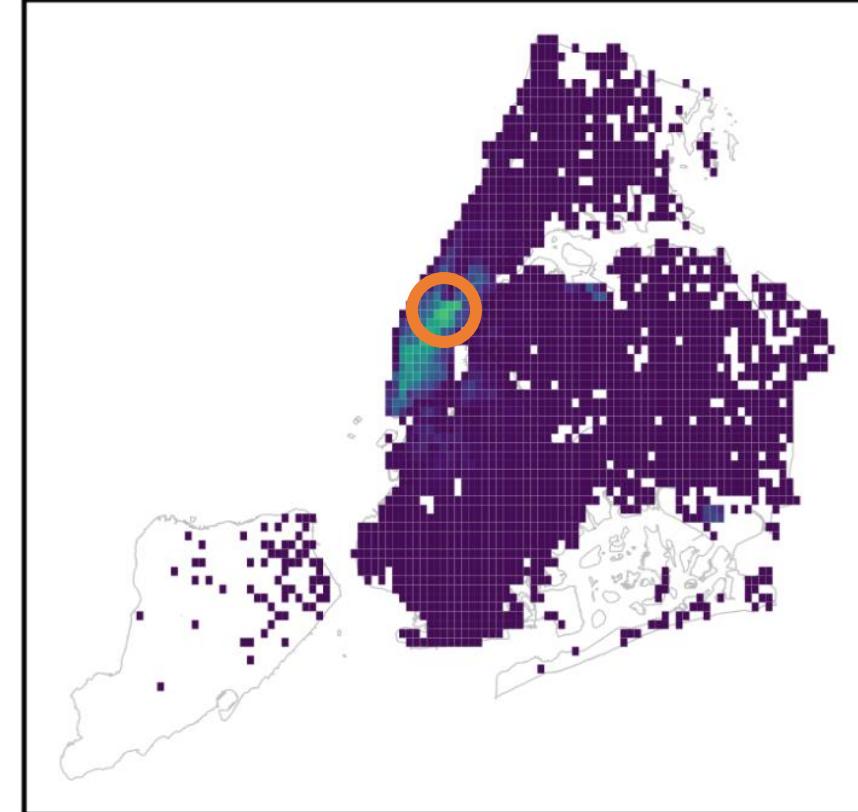


Model Result – Spacial Accuracy

Observe Uber Demand

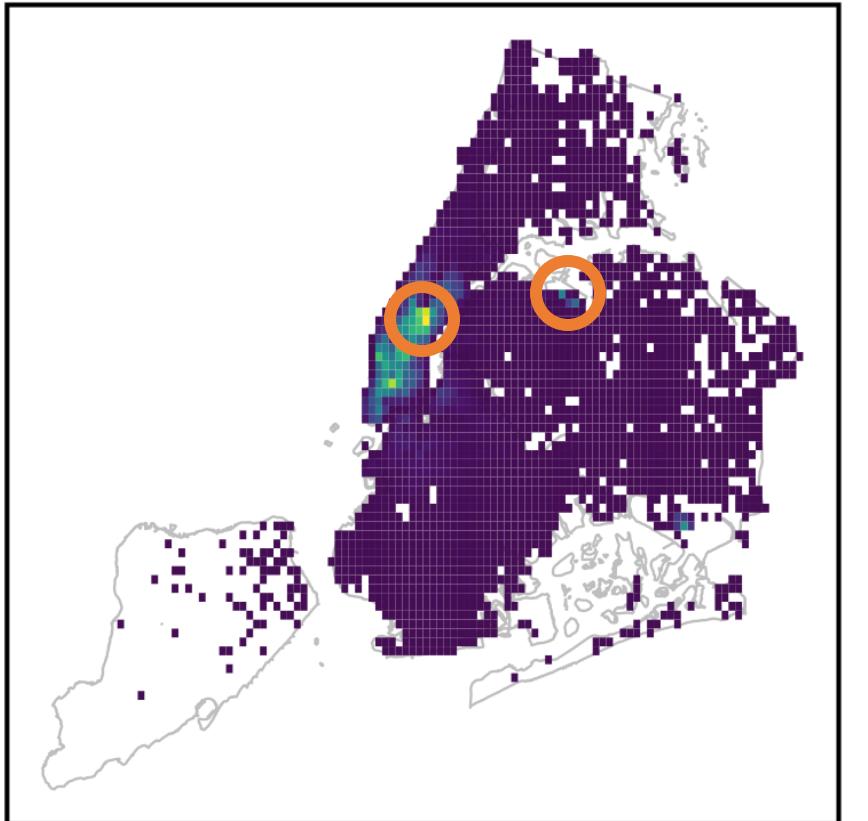


Predicted Uber Demand

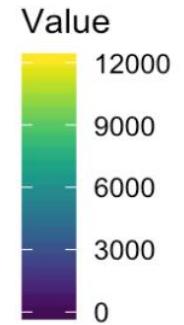
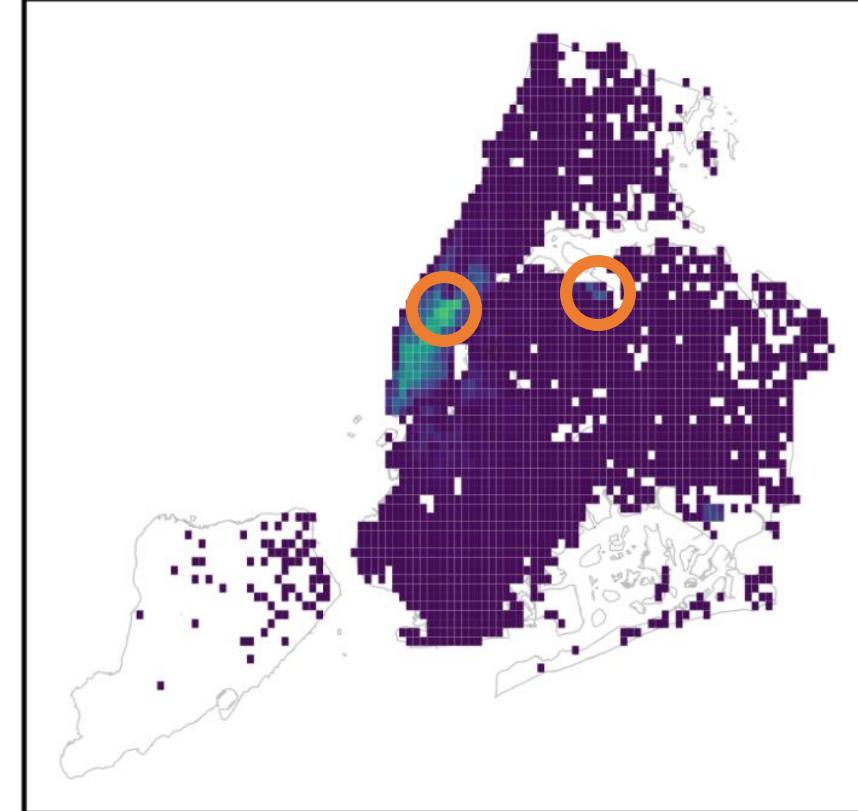


Model Result – Spacial Accuracy

Observe Uber Demand

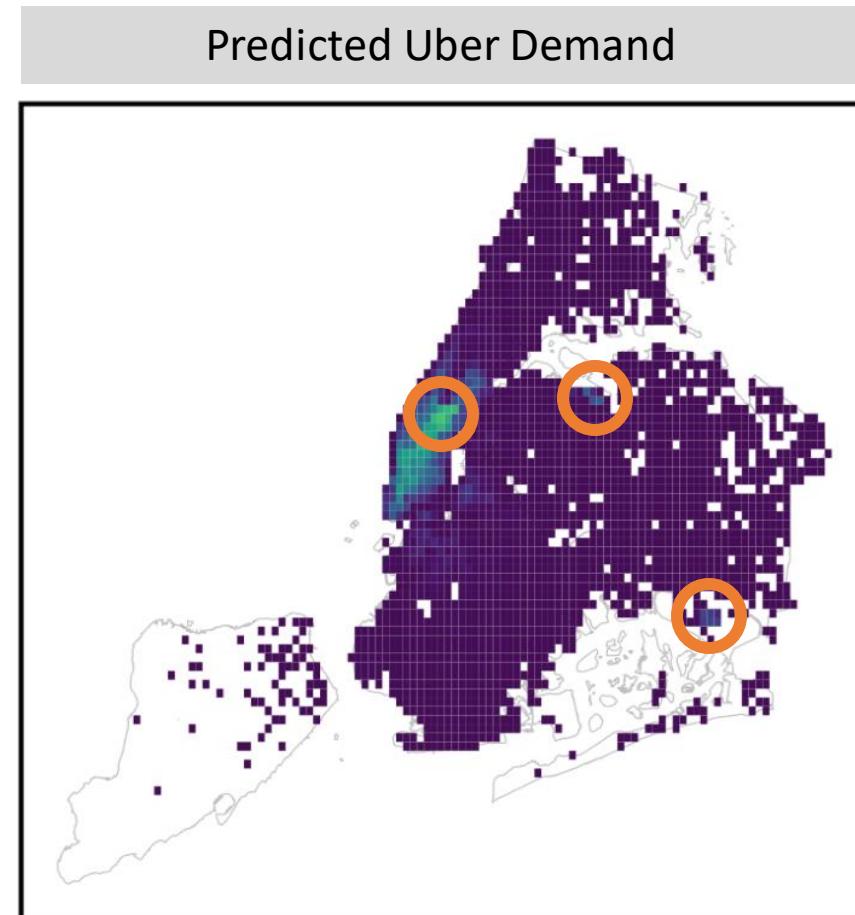
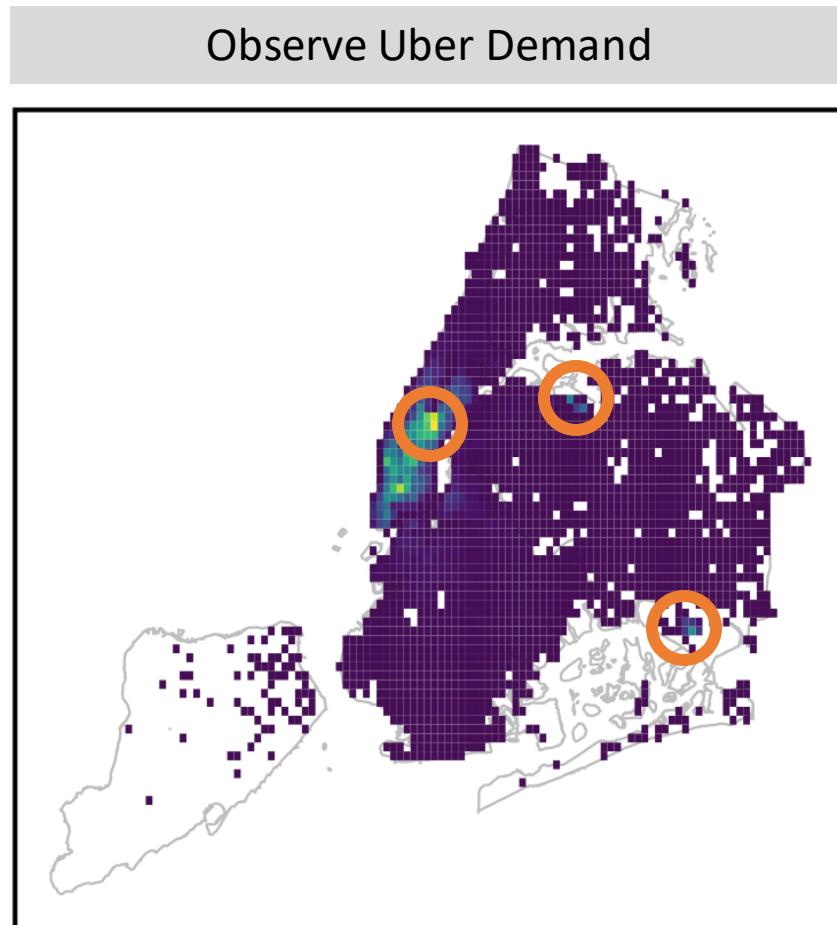


Predicted Uber Demand



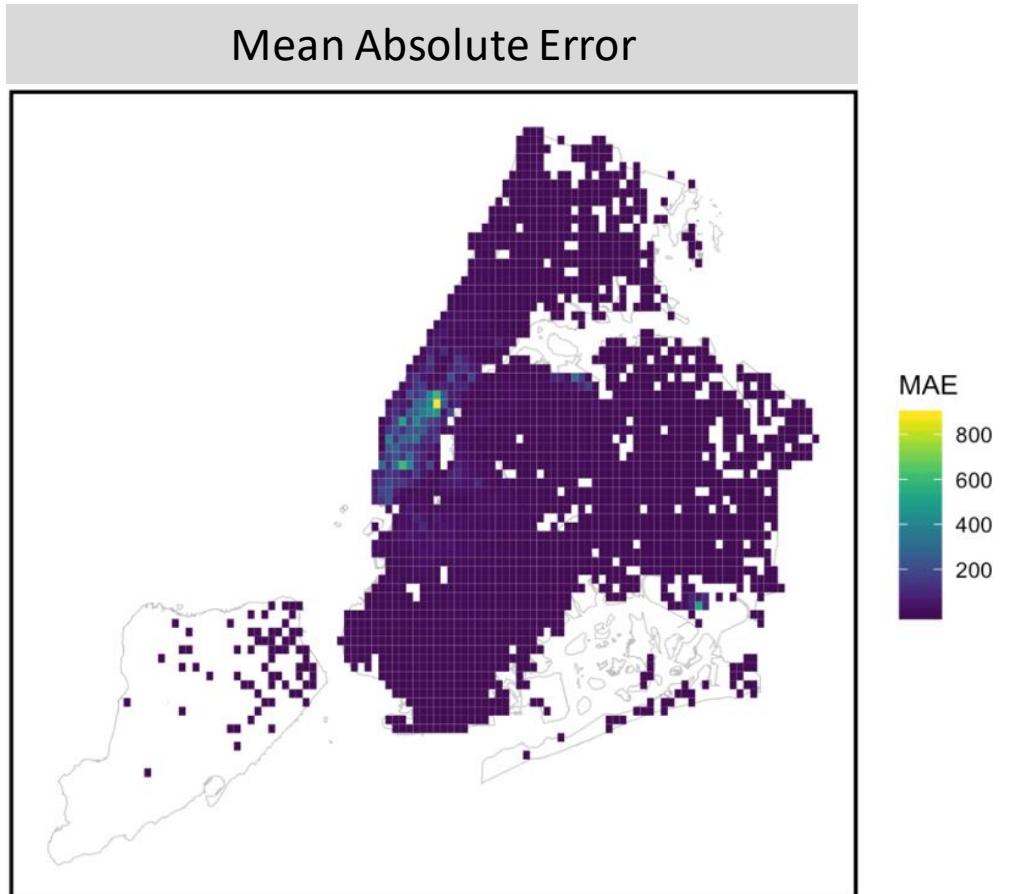
Model Result – Spacial Accuracy

- The model successfully predict the hotspots of Uber demand;
- BUT **underestimate** the popularity of hotspots;



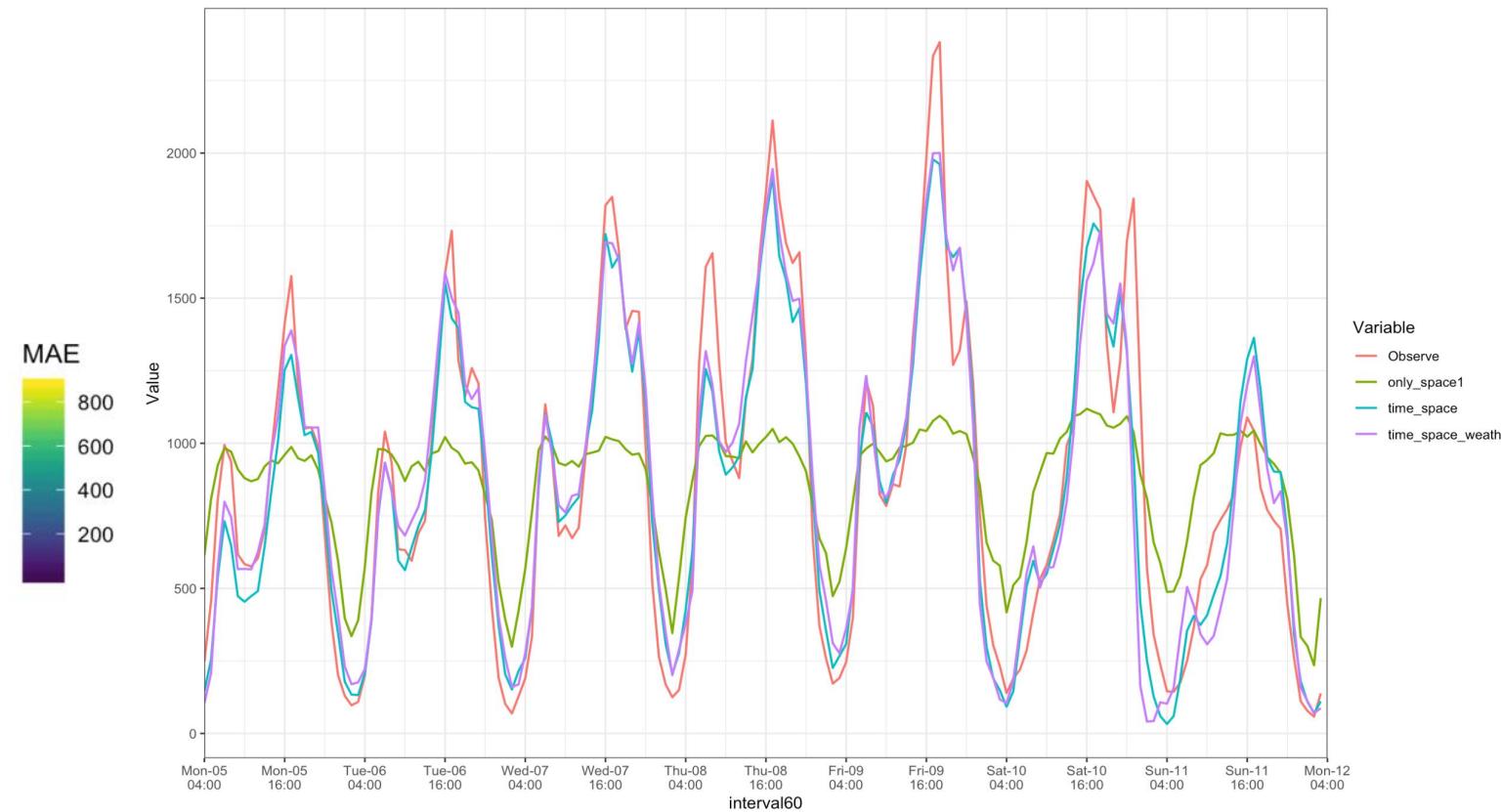
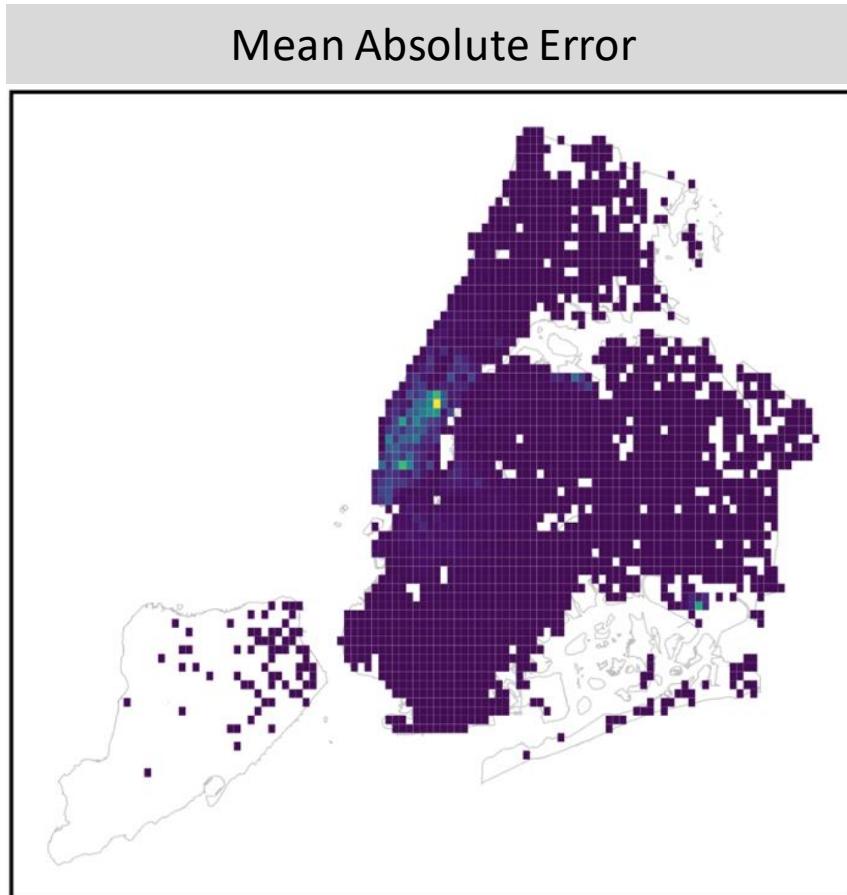
Model Result – Spacial Accuracy

- The model successfully predict the hotspots of Uber demand;
- BUT **underestimate** the popularity of hotspots;
- Mean Absolute Error: for the busiest area, the predicted demand is **800 less** than the observation.



Model Result – Spacial Accuracy

- Temporal Accuracy tell the same story, the model is underperformed in predicting extremely high demand.



Summary

1. Raster-based analysis (although need to **rotate the grids** according to the urban fabric & street later)
2. Uber demand in NYC is clustered (**Moran's I**).
3. Explored the relationship between Uber demand and 3 types of variables (**spacial, temporal and weather**)
4. The model is accurate in general while **underperformed** for high demand prediction.



Thank you!
Q&A